

VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

**Neparametrické a semi-parametrické
modely v analýze přežití**

**Non-parametric and Semi-parametric
Models in Survival Analysis**

2016

Adéla Jelínková

Bibliografický záznam

Autor:	Bc. Adéla Jelínková Fakulta elektrotechniky a informatiky, Vysoká škola báňská – Technická univerzita Ostrava Katedra aplikované matematiky
Název práce:	Neparametrické a semi-parametrické modely v analýze přežití
Studijní program:	Informační a komunikační technologie
Studijní obor:	Výpočetní matematika
Vedoucí práce:	Ing. Martina Litschmannová, Ph.D.
Akademický rok:	2015/16
Počet stran:	ix + 65
Klíčová slova:	Analýza přežití, cenzorování, křivka přežití, Kaplanův-Meierův model, Coxův model, proporcionalita rizik

Bibliographic Entry

Author:	Bc. Adéla Jelínková Faculty of Electrical Engineering and Computer Science, VSB – Technical University of Ostrava Department of Applied Mathematics
Title of Thesis:	Non-parametric and Semi-parametric Models in Survival Analysis
Degree Programme:	Information and Communication Technology
Field of Study:	Computational Mathematics
Supervisor:	Ing. Martina Litschmannová, Ph.D.
Academic Year:	2015/16
Number of Pages:	ix + 65
Keywords:	Survival Analysis, censoring, survival curve, Kaplan-Meier model, Cox model, proportional hazard

Abstrakt

Cílem této diplomové práce je popsat a na reálná data aplikovat neparametrické a semi-parametrické metody analýzy přežití. K dosažení tohoto cíle jsou v práci představeny neparametrické metody, a to Kaplanův-Meierův model, Log-rank test a metoda úmrtnostních tabulek, a dále semi-parametrický Coxův model. Práce se zabývá také regresní diagnostikou, zejména analýzou reziduí, hodnocením vhodnosti modelu a jeho verifikací. Prezentovaná teorie je nakonec aplikována na reálná lékařská data.

Abstract

The aim of the diploma thesis is to describe non-parametric and semi-parametric methods of survival analysis and to test them on the real data. To achieve this, there are presented non-parametric models including Kaplan-Meier model, Log-rank test and life-table method, and semi-parametric Cox model. Furthermore the thesis focuses on regression diagnostic tools, especially residual analysis, a goodness of fit of model and its verification. Eventually, the presented theory is applied to real oncological data.

Zadání diplomové práce

Student: **Bc. Adéla Jelínková**

Studijní program: N2647 Informační a komunikační technologie

Studijní obor: 1103T031 Výpočetní matematika

Téma: **Neparametrické a semi-parametrické modely v analýze přežití**
Non-parametric and semi-parametric models in survival analysis

Jazyk vypracování: čeština

Zásady pro vypracování:

Analýza přežití je souborem statistických metod, které zkoumají dobu do výskytu sledované události, kterou v biomedicíně typicky představuje výskyt onemocnění nebo úmrtí pacienta. Hlavním znakem datových souborů, které zkoumáme metodami analýzy přežití, je fakt, že mohou obsahovat cenzorovaná data, která obsahují pouze částečnou informaci o době do výskytu události.

Cílem práce je popsat, porovnat a aplikovat na aktuální onkologická data poskytnutá onkologickou klinikou, Fakultní nemocnice Ostrava neparametrické a semi-parametrické metody analýzy přežití.

Postup práce:

1. Studium regresních modelů používaných v analýze přežití.
2. Seznámení se s úlohou regresní diagnostiky v hodnocení regresních modelů přežití.
3. Seznámení se s neparametrickými metodami statistické indukce v analýze přežití.
4. Aplikace neparametrických a semi-parametrických metod analýzy přežití na reálná onkologická data.
5. Vyhodnocení analýzy.

Seznam doporučené odborné literatury:

- [1] HOSMER D. W., LEMESHOW S., MAY S., 2008, Applied survival analysis : regression modeling of time-to-event data. 2nd ed. Hoboken, N.J. : Wiley-Interscience, 2008.
- [2] BRENNER H., GEFELLER O., 1996, An alternative approach to monitoring cancer patient survival. Cancer, 78, 2001-2004.

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **Ing. Martina Litschmannová, Ph.D.**

Datum zadání: 01.09.2014

Datum odevzdání: 29.04.2016



Jiří Bouchala

doc. RNDr. Jiří Bouchala, Ph.D.
vedoucí katedry

ay

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Poděkování

Na tomto místě bych chtěla poděkovat zejména své vedoucí diplomové práce, Ing. Martině Litschmannové, Ph.D., za její trpělivost, věnovaný čas, cenné rady a připomínky při vypracování této práce a příležitost, kterou mi poskytla. Dále bych chtěla poděkovat Ing. Kateřině Janurové, Ph.D. za odborné rady. Rovněž bych chtěla poděkovat MUDr. Jakubu Cvekovi, Ph.D. za lékařské konzultace a v neposlední řadě také onkologické klinice Fakultní nemocnice Ostrava za poskytnutá data, bez kterých by tato diplomová práce nemohla vzniknout.

Prohlášení

Prohlašuji, že jsem svoji diplomovou práci vypracovala samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Ostrava 28. dubna 2016

.....
Bc. Adéla Jelínková

Obsah

Úvod	1
Přehled použitého značení	2
Kapitola 1. Analýza přežití	4
1.1 Cenzorování	4
1.2 Důležité pojmy	6
Kapitola 2. Neparametrické metody	8
2.1 Kaplanův-Meierův model	8
2.1.1 Greenwoodova formule	10
2.1.2 Intervaly spolehlivosti	11
2.1.3 Medián doby přežití	12
2.2 Porovnání dvou nebo více křivek přežití	12
2.3 Metoda úmrtnostních tabulek	13
Kapitola 3. Semi-parametrické metody	15
3.1 Coxův model	15
3.1.1 Poměr rizik	16
3.1.2 Odhad regresních koeficientů	16
3.1.3 Interval spolehlivosti pro regresní koeficienty	18
3.1.4 Testy hypotéz pro regresní koeficienty	18
3.2 Sestavení Coxova modelu	20
Kapitola 4. Regresní diagnostika	21
4.1 Analýza reziduí	21
4.1.1 Martingale rezidua	22
4.1.2 Deviační rezidua	22
4.1.3 Skóre rezidua	24
4.1.4 Schoenfeldova rezidua	25
4.2 Hodnocení vhodnosti modelu	26
4.2.1 Grønnesbyho a Borganův test	26
4.2.2 Akaikeho informační kritérium	27
4.3 Verifikace modelu	27
4.3.1 Hodnocení proporcionality rizik	27

4.3.2	Identifikace vlivných pozorování	29
4.3.3	Identifikace odlehlých hodnot	29
Kapitola 5.	Analýza reálných dat	30
5.1	Neparametrická analýza	33
5.1.1	Log-rank test	33
5.2	Semi-parametrická analýza	47
5.2.1	Coxův model	47
5.3	Aplikace regresní diagnostiky	52
5.3.1	Analýza reziduí	52
5.3.2	Hodnocení vhodnosti modelu	58
5.3.3	Hodnocení proporcionality rizik	58
5.4	Interpretace výsledného modelu	60
Závěr	63
Seznam použité literatury	64

Úvod

V dnešní době je analýza přežití součástí mnoha odvětví. Příklady jejího použití bychom mohli najít v medicíně, biologii, ekonomii, demografii, technických oborech, atd. Její zásadní charakteristikou je schopnost pracovat s daty v cenzorovaném tvaru. Hlavním rysem, který odlišuje data o přežití od ostatních typů dat (např. od klasického vnímání úmrtnosti jako podílu zemřelých pacientů) je jejich časová složka. Data o přežití totiž odráží nejen informaci o počtu, respektive podílu sledovaných událostí, ale zároveň nás informují, kdy k dané události došlo.

Analýza přežití zahrnuje matematicko-statistické metody pro hodnocení doby do výskytu dané události, které se však liší od standardního statistického hodnocení, neboť data popisující dobu do výskytu události mají řadu specifik.

Cílem naší práce je seznámit se s metodami analýzy přežití a ty poté aplikovat na reálná data. Máme k dispozici onkologická data, která poskytla onkologická klinika Fakultní nemocnice Ostrava. Sběr dat proběhl v období 2008 až 2013.

V první kapitole se podíváme na analýzu přežití obecně a definujeme základní pojmy, v druhé kapitole popíšeme neparametrické metody, jako jsou Kaplanův-Meierův model, Log-rank test a metoda úmrtnostních tabulek. Ve třetí kapitole se budeme zabývat semi-parametrickými metodami analýzy přežití, zejména Coxovým modelem proporcionálních rizik. Čtvrtá kapitole bude zaměřena na regresní diagnostiku, která je neméně důležitá jako samotná analýza dat. A nakonec v páté kapitole aplikujeme získané poznatky na konkrétní data a vyhodnotíme výsledky.

Veškeré statistické výpočty, modelování a vykreslování obrázků provedeme ve statistickém softwaru R Studio s využitím balíčku pro analýzu přežití – `survival`.

Přehled použitého značení

AIC	Akaikeho informační kritérium
β_i	regresní koeficienty
c_i	počet cenzorování v čase $t_i \leq t$
C	doba cenzorování
d_i	počet sledovaných událostí v čase $t_i \leq t$
$D_i(t)$	deviační reziduum
δ	indikátor cenzury
\hat{E}_i	očekávaný počet událostí
$E(X)$	střední hodnota
$f(t)$	hustota pravděpodobnosti
$F(t)$	distribuční funkce
H_0	nulová hypotéza
H_1	alternativní hypotéza
$h(t)$	riziková funkce
$H(t)$	kumulativní riziková funkce
$h(t, X)$	Coxův model proporcionálních rizik
HR	poměr rizik
$I(\beta)$	informační matice
IS	interval spolehlivosti
KM	Kaplanův-Meierův (model)
$l(\beta)$	logaritmus parciální věrohodnostní funkce
$L(\beta)$	parciální věrohodnostní funkce
L_{im}	skóre reziduum
M_i	množina subjektů v riziku v čase t_i
$M_i(t)$	martingale reziduum
R_i	počet pacientů v riziku v době t_i
$S(t)$	funkce přežití
$SE(X)$	standardní chyba
R_{im}	Schoendfeldovo reziduum
T	doba přežití

TPVP	test parciálním věrohodnostním poměrem
$U(\beta)$	skórový vektor
$var(X)$	rozptyl
$var(\mathbf{X})$	kovarianční matice
\mathbf{X}'	transponovaná matice \mathbf{X}
\mathbf{X}^{-1}	inverzní matice k matici \mathbf{X}
z	testová statistika Waldova testu
$z_{1-\frac{\alpha}{2}}$	$100(1 - \frac{\alpha}{2})\%$ kvantil standardizovaného normálního rozdělení
ZN	zhoubný novotvar
$\frac{\partial}{\partial x}$	parciální derivace podle x

Kapitola 1

Analýza přežití

Analýza přežití byla ve svých počátcích využívána pro odhadování očekávané doby do úmrtí pacienta se smrtelným onemocněním. V dnešní době se tato statistická metoda užívá v mnoha dalších odvětvích, avšak název metody zůstal zachován. V našem případě budeme používat tuto metodu v originálním významu, tedy pro analýzu dat z onkologického ústavu.

Analýza přežití se obecně zabývá hodnocením doby od počáteční události do výskytu jedné nebo více určitých událostí (*time-to-event analysis*), (Clark, Bradburn, Love a Altman, 2003). Počáteční událostí může být narození, začátek onemocnění nebo například začátek léčby. Druhou, tzv. sledovanou událostí (*event of interest*) je například úmrtí, uzdravení či remise onemocnění (neboli vymizení příznaků a projevů onemocnění, nikoliv nemoci samotné). Zkoumanou náhodnou veličinou je doba mezi těmito dvěma okamžiky, tzv. doba přežití (*survival time*) neboli doba do výskytu sledované události. Tuto veličinu označujeme písmenem T . Sledovanou proměnnou nejčastěji bývá doba do úmrtí pacienta.

Touto statistickou metodou dokážeme řešit následující problémy:

1. bodový a intervalový odhad dob přežití, odhadování pravděpodobnostního rozdělení doby přežití;
2. porovnávání dob přežití mezi několika různými skupinami osob (např. muži a ženy), odhadování vlivu léčby na dobu přežití;
3. modelování a hodnocení vlivu různých potenciálních rizikových faktorů na přežití (např. věk při provedení operace, velikost dávky léku).

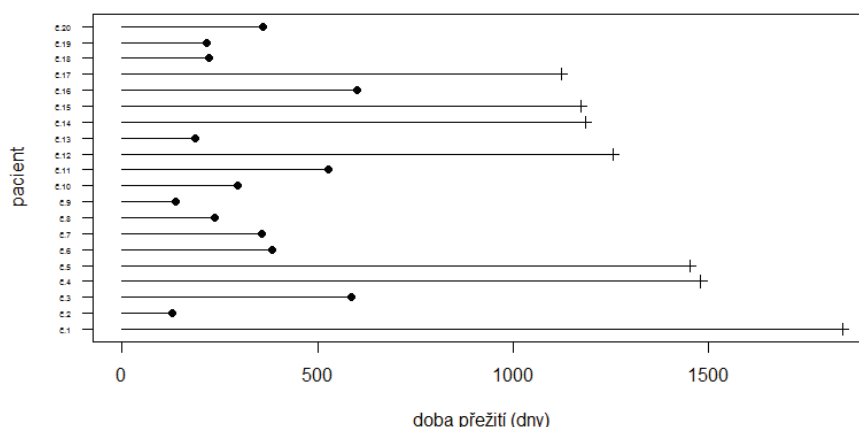
1.1 Cenzorování

Typickým rysem analýzy přežití je to, že pracuje s neúplnými daty. Studie bývá v praxi většinou ukončena dříve, než se u všech pacientů vyskytne sledovaná událost. To znamená, že nemáme všechna data kompletní a skutečná doba do výskytu sledované události je v těchto případech neznámá. Toto se nazývá **cenzorování** a data, o kterých víme, že nereprezentují reálné doby do výskytu sledovaných událostí, nazýváme **cenzorovaná data**.

(Zvárová & Malý, 2003).

Cenzorování může nastat z různých důvodů. Například pacient přestane spolupracovat, odstěhuje se, nebo u něj dojde k události z jiných než sledovaných příčin. Doba přežití je tedy neznámá, avšak víme, že je delší než doba pozorování. Dalším důvodem cenzorování je předem stanovená doba klinické studie. Zde jsou známy doby přežití jen pro pacienty, u kterých událost nastala v průběhu této studie. O ostatních pak víme pouze to, že jejich doba přežití je delší než doba pozorování do ukončení sledování.

Cenzorování z výše uvedených důvodů se nazývá **cenzorováním zprava** a v analýze přežití se vyskytuje nejčastěji.



Obrázek 1.1: Cenzorování

Dalším typem je **cenzorování zleva**. Tato situace nastane tehdy, když je pacient do studie zařazen až poté, co se u něj sledovaná událost vyskytla. Třetím typem cenzorování je **intervalové**, jenž je kombinací cenzorování zprava a zleva.

Dále budeme pracovat s daty, jež jsou cenzorována zprava. Tento typ cenzorování se objevuje na našich reálných datech. To můžeme vidět na Obrázku 1.1, kde je znázorněna situace pro prvních 20 pacientů studie. Událost (úmrť) je vyznačeno „●“ a „+“ značí cenzorovaný záznam.

Data o přežití můžeme nejlépe vyjádřit dvojicí proměnných (T, C) , kde T je náhodná veličina, která popisuje dobu sledování (doba přežití nebo v případě cenzorování doba strávená ve studii) a C je indikátor sledované události, resp. indikátor cenzorování, a nabývá hodnot $C = 1$, pokud událost nastala a $C = 0$ v případě, kdy je doba přežití cenzorována.

1.2 Důležité pojmy

K popisu pravděpodobnostního rozdělení náhodných veličin běžně používáme **distribuční funkci** F a **hustotu pravděpodobnosti** f . V analýze přežití distribuční funkce udává pravděpodobnost, že doba přežití T je kratší nebo rovna t , $t \geq 0$, tj.

$$F(t) = P(T \leq t). \quad (1.1)$$

Hustotu pravděpodobnosti získáme derivací distribuční funkce

$$f(t) = \frac{d}{dt}F(t). \quad (1.2)$$

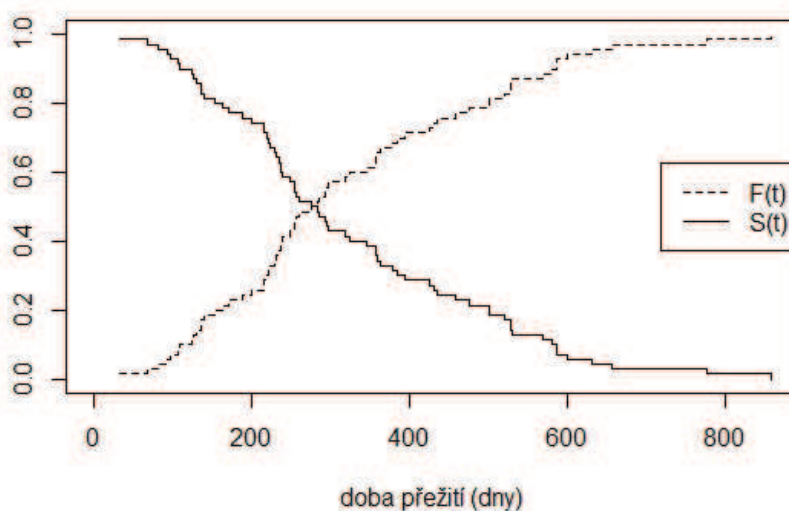
Funkce přežití

V analýze přežití zavádíme navíc tzv. **funkci přežití** (*survival function*)

$$S(t) = 1 - F(t) = P(T > t). \quad (1.3)$$

Tato funkce vyjadřuje pravděpodobnost, že v daném časovém intervalu $(0, t)$ nedojde ke sledované události, tedy že doba přežití T překročí hodnotu t .

Na Obrázku 1.2 je pro srovnání znázorněn průběh distribuční funkce a funkce přežití.



Obrázek 1.2: Distribuční funkce $F(t)$ a funkce přežití $S(t)$

Riziková funkce

Jestliže je distribuční funkce F spojitá, pak můžeme definovat tzv. **rizikovou funkci** $h(t)$ (*hazard function*). Ta představuje derivaci podmíněné pravděpodobnosti výskytu sledované

události za předpokladu, že do času t tato událost ještě nenastala. Jedná se o „okamžitou intenzitu“ sledované události během nekonečně malého časového intervalu délky Δt (Rodríguez, 2007).

$$h(t) = \lim_{\Delta t \rightarrow 0+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}, \text{ pro } F(t) \neq 1 \quad (1.4)$$

Kumulativní riziková funkce

Integrací předchozího vztahu získáme tzv. **kumulativní rizikovou funkci** (*cumulative hazard function*)

$$H(t) = \int_0^t h(u) du = -\ln S(t) + \ln S(0) = -\ln S(t). \quad (1.5)$$

Odtud po úpravě dostaneme

$$S(t) = e^{-H(t)}. \quad (1.6)$$

Mezi výše uvedenými pojmy tedy existují jednoduché převodní vztahy. V Tabulce 1.1 jsou uvedeny hlavní z nich.

	$F(t)$	$S(t)$	$f(t)$	$h(t)$
$F(t)$	—	$1 - S(t)$	$\int_0^t f(x) dx$	$1 - \exp \left[- \int_0^t h(x) dx \right]$
$S(t)$	$1 - F(t)$	—	$1 - \int_0^t f(x) dx$	$\exp \left[- \int_0^t h(x) dx \right]$
$f(t)$	$\frac{dF(t)}{dt}$	$-\frac{dS(t)}{dt}$	—	$h(t) \exp \left[- \int_0^t h(x) dx \right]$
$h(t)$	$\frac{\frac{dF(t)}{dt}}{1 - F(t)}$	$-\frac{d(\ln(S(t)))}{dt} = \frac{-\frac{dS(t)}{dt}}{S(t)}$	$\frac{f(t)}{1 - \int_0^t f(x) dx}$	—

Tabulka 1.1: Převodní vztahy

Kapitola 2

Neparametrické metody

Metody v analýze přežití dělíme podle toho, co víme o pravděpodobnostním rozdělení náhodné veličiny T (doba přežití) do tří základních skupin na:

1. parametrické,
2. neparametrické,
3. semi-parametrické.

Pro použití **parametrických metod** musíme znát pravděpodobnostní rozdělení doby přežití. Tyto metody jsou obecně nejjednodušší a zároveň podávají nejpřesnější informace. Avšak v praxi většinou rozdělení pravděpodobnosti doby přežití dopředu neznáme, tedy tyto metody nelze obvykle použít.

Naproti tomu **neparametrické metody** lze použít vždy, protože tyto metody vycházejí jen z naměřených hodnot. Jsou obecnější než parametrické metody. Nejpoužívanější mezi neparametrickými metodami jsou Kaplanův-Meierův odhad a metoda úmrtnostních tabulek.

Posledním typem jsou **semi-parametrické metody**, které jsou přechodem mezi parametrickými a neparametrickými metodami. Zde nepotřebujeme konkrétní pravděpodobnostní rozdělení náhodné veličiny T , ale využíváme předpoklady ohledně vysvětlujících parametrů modelu. Nejpoužívanější metodou v této skupině je Coxův model proporcionálních rizik.

V této práci se budeme podrobněji zabývat neparametrickými a semi-parametrickými metodami, konkrétně Kaplan-Meierovým a Coxovým modelem. Tyto dva modely také aplikujeme na konkrétní reálná onkologická data.

2.1 Kaplanův-Meierův model

Jak už bylo řečeno, **Kaplanův-Meierův model** (dále KM model) je nejpoužívanější z neparametrických metod. Využívá se pro odhad funkce přežití na základě pozorovaných dat

(Kaplan a Meier, 1958). Pro tento odhad, též nazývaný „*product-limit*“ odhad, požadujeme pouze jeden předpoklad – aby doba cenzorování byla nezávislá na době výskytu události. To znamená: aby byl pacient v době t bez sledované události, nesmí se u něj událost vyskytnout v žádné době t^* takové, pro kterou platí, že $t^* < t$. Pro odhad pravděpodobnosti, že u daného pacienta se do doby t nevyskytne sledovaná událost, je potřeba odhadnout odpovídající pravděpodobnosti pro všechny doby t^* , které předcházely době t . Budeme předpokládat n různých dob přežití takových, že $t_1 < t_2 < \dots < t_n < t$. Pak můžeme vyjádřit pravděpodobnost přežití až do doby t vztahem

$$\begin{aligned} S(t) &= P(T > t) = P((T > t_1) \cap (T > t_2) \cap \dots \cap (T > t_{n-1}) \cap (T > t_n)) = \\ &= P(T > t_1)P(T > t_2 | T > t_1) \cdot \dots \cdot P(T > t_{n-1} | T > t_{n-2})P(T > t_n | T > t_{n-1}). \end{aligned} \quad (2.1)$$

Vzhledem k tomu, že se jedná o neparametrickou metodu, máme k dispozici jako vstupní informace pouze pozorované hodnoty dob přežití. Pravděpodobnost přežití doby t_i tedy musíme vyjádřit jen pomocí údajů o počtu úmrtí v daném čase a počtu pacientů v riziku:

$$\hat{P}(T > t_i | T > t_{i-1}) = \hat{p}_i = \frac{R_i - d_i}{R_i} = 1 - \frac{d_i}{R_i}, \quad (2.2)$$

kde d_i je počet sledovaných událostí v čase $t_i \leq t$ a R_i je počet pacientů v riziku v době t_i , což je počet pacientů, kteří přežili dobu t_{i-1} . Funkci přežití pak můžeme odhadnout vztahem

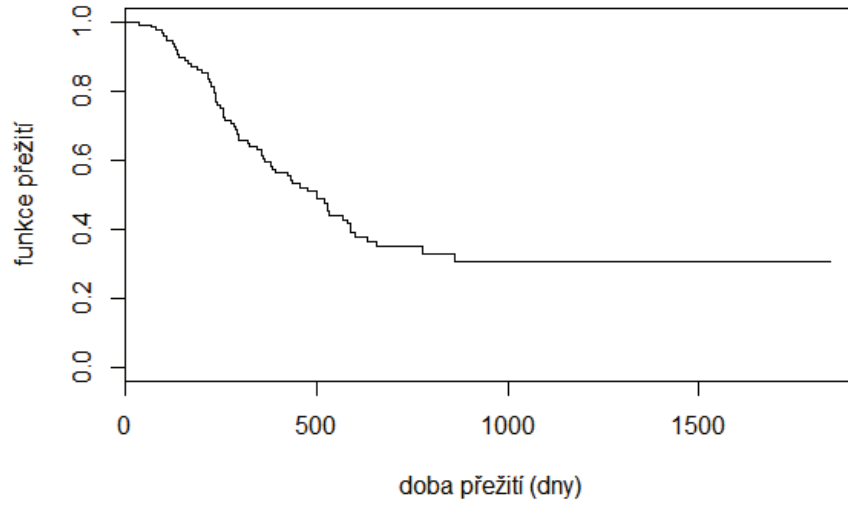
$$\hat{S}(t) = \hat{p}_1 \cdot \hat{p}_2 \cdot \dots \cdot \hat{p}_n = \frac{R_1 - d_1}{R_1} \cdot \frac{R_2 - d_2}{R_2} \cdot \dots \cdot \frac{R_{n-1} - d_{n-1}}{R_{n-1}} \cdot \frac{R_n - d_n}{R_n}. \quad (2.3)$$

Při odhadování pravděpodobností přežití jednotlivých dob t_i však musíme zohlednit cenzorování, protože s cenzorovanými dobami přežití nemůžeme pracovat stejně jako s kompletním pozorováním. Tato cenzorovaná data nepřispívají k d_i , ale nelze je při hodnocení vynechat. Kaplanův-Meierův odhad pracuje s cenzorováním tak, že tyto pacienty nezahrnuje do skupiny pacientů v riziku po zaznamenaném čase cenzorování. To znamená, že jestliže je doba t cenzorovaná a platí, že $t_i < t < t_{i+1}$, potom se daný pacient v době t_i započítává do skupiny subjektů v riziku R_i , ale do následující pozorované doby výskytu sledované události t_{i+1} se už do skupiny v riziku R_{i+1} nezahrnuje. Ze vztahu (2.3) tedy dostáváme výsledný vzorec pro Kaplanův-Meierův odhad funkce přežití:

$$\hat{S}(t) = \prod_{t_i \leq t} \hat{p}_i = \prod_{t_i \leq t} \left(1 - \frac{d_i}{R_i}\right). \quad (2.4)$$

Součin v rovnici (2.4) vlastně počítáme jen přes necenzorované doby přežití. Avšak pro cenzorované doby je počet sledovaných událostí $d_i = 0$, tzn. cenzorované doby přežití k tomuto odhadu přispívají pouze přes hodnotu počtu pacientů v riziku R_i . Tento součin lze tedy teoreticky definovat přes všechna pozorování.

Na Obrázku 2.1 si můžeme prohlédnout, jak tento odhad funkce přežití bude vypadat pro naše konkrétní onkologická data.



Obrázek 2.1: Kaplanův-Meierův odhad funkce přežití

2.1.1 Greenwoodova formule

Podívejme se nyní na rozptyl Kaplanova-Meierova odhadu funkce přežití. KM odhad je dán jako součin \hat{p}_i , proto bude lepší jej nejprve zlogaritmovat, čímž dostaneme součet logaritmů jednotlivých odhadů \hat{p}_i (Hosmer, Lemeshow a May, 2008):

$$\text{var}(\ln \hat{S}(t)) = \text{var}(\ln \prod_{t_i \leq t} \hat{p}_i) = \text{var}(\sum_{t_i \leq t} \ln \hat{p}_i) = \sum_{t_i \leq t} \text{var}(\ln \hat{p}_i). \quad (2.5)$$

Pro výpočet rozptylu logaritmu \hat{p}_i využijeme maximálně věrohodný odhad pravděpodobnosti p_i , tedy $1 - \frac{d_i}{R_i}$, a tzv. delta metodu (*delta method*, viz např. Hosmer, Lemeshow a May, 2008):

$$\text{var}(\ln \hat{p}_i) = \left(\frac{\partial}{\partial p_i} \ln p_i \Big|_{p_i = \hat{p}_i} \right)^2 \text{var}(\hat{p}_i) = \left(\frac{1}{\hat{p}_i} \right)^2 \frac{\hat{p}_i(1 - \hat{p}_i)}{R_i} = \frac{1 - \hat{p}_i}{R_i \hat{p}_i} = \frac{d_i}{R_i(R_i - d_i)}. \quad (2.6)$$

Z rovnic (2.5), (2.6) dostaneme

$$\text{var}(\ln \hat{S}(t)) = \sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}, \quad (2.7)$$

použijeme znovu delta metodu

$$\text{var}(\ln \hat{S}(t)) = \left(\frac{\partial}{\partial \hat{S}(t)} \ln \hat{S}(t) \right)^2 \text{var}(\hat{S}(t)) = \left(\frac{1}{\hat{S}(t)} \right)^2 \text{var}(\hat{S}(t)) \quad (2.8)$$

a odsud už snadno dostáváme vzorec pro výpočet rozptylu KM odhadu funkce přežití, tzv. **Greenwoodovu formuli** (Greenwood, 1926)

$$\text{var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}. \quad (2.9)$$

Greenwoodova formule je nejběžněji používaný vzorec pro odhad rozptylu Kaplanovy-Meierovy funkce přežití a také jej můžeme nalézt implementovaný téměř ve všech softwarech, které nám umožňují řešit analýzu přežití. Nicméně, existuje mnoho dalších, jako příklad zde uvedeme ještě Aalenův-Johansenův odhad rozptylu KM funkce přežití (Andersen, Borgan, Gill a Keiding, 2012):

$$\text{var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{t_i \leq t} \frac{d_i}{R_i^2}. \quad (2.10)$$

2.1.2 Intervaly spolehlivosti

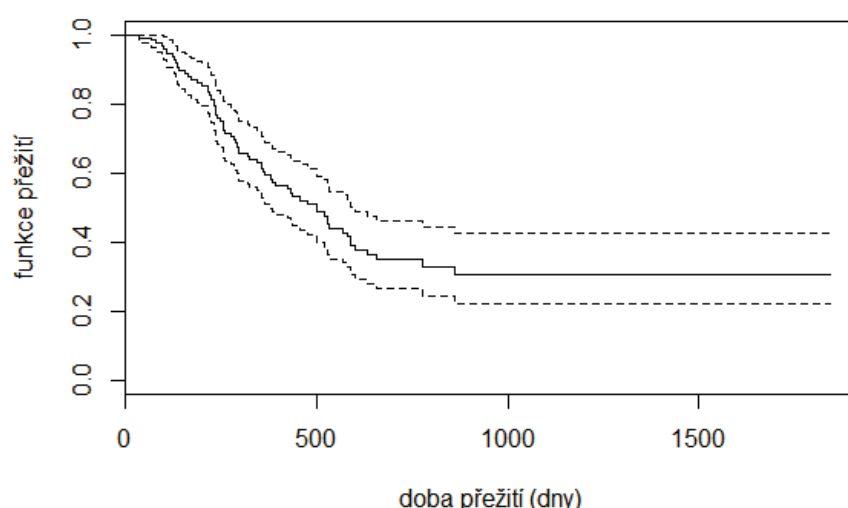
S využitím Greenwoodovy formule můžeme sestavit $100(1 - \alpha)\%$ **interval spolehlivosti** pro Kaplanův-Meierův odhad funkce přežití $\hat{S}(t)$ v čase t tak, že doby přežití, za předpokladu, že máme dostatečně mnoho subjektů, aproximujeme normálním rozdělením, (Hosmer, Lemeshow a May, 2008):

$$\left[\hat{S}(t) - z_{1-\alpha/2} \sqrt{\text{var}(\hat{S}(t))}; \hat{S}(t) + z_{1-\alpha/2} \sqrt{\text{var}(\hat{S}(t))} \right], \quad (2.11)$$

kde $z_{1-\alpha/2}$ je $100(1 - \alpha/2)\%$ kvantil standardizovaného normálního rozdělení.

Tento postup je nejpoužívanější, kvůli výpočetní jednoduchosti a dostupnosti. Má však i velký nedostatek – symetrii. V blízkosti hodnot 1 a 0 je totiž symetrie intervalu spolehlivosti pro odhad funkce přežití nevhodná, protože připouští hodnoty přežití větší než 1 nebo naopak menší než 0. Proto se někdy používá škálování.

Na Obrázku 2.2 vidíme, jak tyto intervaly spolehlivosti vypadají pro naše reálná data.



Obrázek 2.2: Kaplanův-Meierův odhad funkce přežití spolu s intervaly spolehlivosti

2.1.3 Medián doby přežití

Medián doby přežití (*median survival time*), definujeme jako dobu $t_{0,5}$, kdy má funkce přežití hodnotu 0,5 (Collett, 2003). Platí

$$S(t_{0,5}) = 0,5.$$

Medián přežití je doba, ve které je pravděpodobnost přežití 50%, tedy doba, které se dožije polovina sledovaných pacientů.

Analogicky můžeme definovat i další kvantily. $100p\%$ kvantil náhodné veličiny, je definován jako doba t_p , pro kterou platí

$$S(t_p) = 1 - p.$$

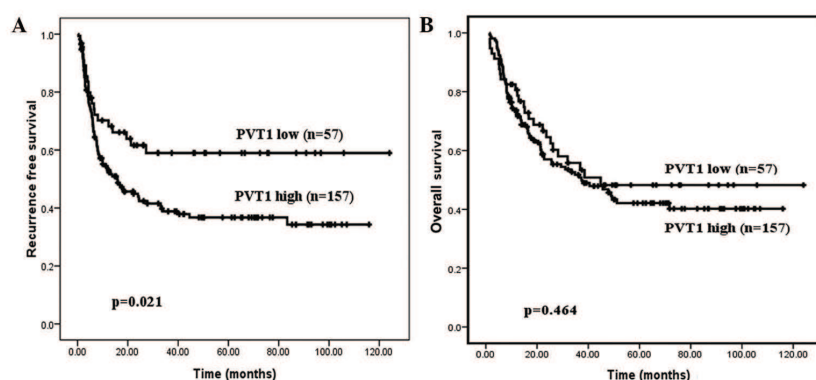
Jako příklad uvádíme 25% a 75% kvantily doby přežití, které jsou definovány jako doby $t_{0,25}$ a $t_{0,75}$, pro které platí

$$S(t_{0,25}) = 0,75 \text{ a } S(t_{0,75}) = 0,25.$$

2.2 Porovnání dvou nebo více křivek přežití

V této podkapitole se stručně podíváme na **log-rank test**, neparametrickou metodu pro porovnávání dvou nebo více křivek přežití (nebo rizikových funkcí) pro cenzorovaná data. Zpravidla platí, že když křivka přežití, která reprezentuje jednu skupinu subjektů leží zcela nad křivkou přežití reprezentující druhou skupinu, potom podíl subjektů z první skupiny, u kterých odhadujeme, že jsou naživu, je v každém časovém okamžiku vyšší než podíl takovýchto subjektů v druhé skupině. Pomocí log-rank testu zjistíme, jestli je tento rozdíl statisticky významný.

Na Obrázku 2.3 si můžeme prohlédnout dva grafy. V levém grafu jedna Kaplanova-Meierova křivka přežití leží celá nad KM křivkou druhé skupiny a log-rank test s p-hodnotou 0,021 potvrdil, že křivky se liší. Vpravo se křivky kříží a p-hodnota log-rank testu je rovna 0,464, tedy dostáváme výsledek, že křivky se statisticky významně neliší.



Obrázek 2.3: Porovnání dvou křivek přežití pomocí log-rank testu (Ding, Yang, Lv a další, 2015)

Tato metoda tedy spočívá v tom, že testujeme nulovou hypotézu o shodnosti dvou nebo více křivek přežití proti alternativní hypotéze, která tvrdí, že mezi křivkami je statisticky významný rozdíl.

Za platnosti nulové hypotézy má testovací statistika Q pro dvě křivky přežití asymptoticky χ^2 –rozdělení s jedním stupněm volnosti (Clark, Bradburn, Love a Altman, 2003):

$$Q = \frac{(U - E(U))^2}{\text{Var}(U)} \sim \chi_1^2,$$

kde

$$\begin{aligned} U &= \sum_{i=1}^I (d_{1i} - E(d_{1i})) = \sum_{i=1}^I (d_{1i} - R_{1i} \frac{d_i}{R_i}), \\ E(U) &= 0, \\ \text{Var}(U) &= \sum_{i=1}^I \text{Var}(d_{1i}) = \sum_{i=1}^I \frac{R_{1i} R_{2i} (R_i - d_i) d_i}{R_i^2 (R_i - 1)}. \end{aligned}$$

2.3 Metoda úmrtnostních tabulek

Další z neparametrických metod pro porovnávání křivek přežití je **metoda úmrtnostních tabulek** (*life-table method*). Příklad úmrtnostní tabulky si můžeme prohlédnout na Obrázku 2.4 na následující straně.

Tato metoda je podobná KM metodě. Avšak zatímco délka intervalů v Kaplanově-Meierově metodě je závislá na časové jednotce měření a může být blízká nule, metoda úmrtnostních tabulek celkovou dobu pozorování dělí na intervaly, jejichž délka je pevně dána.

Celková doba pozorování je tedy rozdělena na k intervalů s indexy $j = 1, 2, \dots, k$. Do úmrtnostní tabulky se pro každý z těchto intervalů zaznamenávají údaje, které jsou potřebné pro odhad pravděpodobnosti přežití (Marubini & Valsecchi, 2004):

$$S_j = 1 - \frac{d_j}{R_j - \frac{1}{2}c_j}, \quad (2.12)$$

kde S_j je podmíněná pravděpodobnost přežití pro j -tý interval, d_j značí počet pozorovaných událostí během j -tého intervalu, R_j je počet subjektů v riziku na počátku j -tého intervalu a c_j představuje počet cenzorovaných dob během j -tého intervalu.

Postupným násobením podmíněných pravděpodobností přežití pro předešlé intervaly získáme kumulativní pravděpodobnost přežití od počátku pozorování do konce j -tého intervalu (Holčík, Komenda & kol., 2015):

$$S(j) = \prod_{i=1}^j S_i = \prod_{i=1}^j \left(1 - \frac{d_i}{R_i - \frac{1}{2}c_i} \right). \quad (2.13)$$

Time	At risk	Dead	Censored	S	SE(S)	H	SE(H)
142	22	1	0	0.954545	0.044409	0.04652	0.046524
157	21	1	0	0.909091	0.061291	0.09531	0.06742
163	20	1	0	0.863636	0.073165	0.146603	0.084717
198	19	1	0	0.818182	0.08223	0.200671	0.100504
204	18	0	1	0.818182	0.08223	0.200671	0.100504
205	17	1	0	0.770053	0.090387	0.261295	0.117378
232	16	3	0	0.625668	0.105069	0.468935	0.16793
233	13	4	0	0.433155	0.108192	0.836659	0.249777
239	9	1	0	0.385027	0.106338	0.954442	0.276184
240	8	1	0	0.336898	0.103365	1.087974	0.306814
261	7	1	0	0.28877	0.099172	1.242125	0.34343
280	6	2	0	0.192513	0.086369	1.64759	0.44864
295	4	2	0	0.096257	0.064663	2.340737	0.671772
323	2	1	0	0.048128	0.046941	3.033884	0.975335
344	1	0	1	0.048128	0.046941	3.033884	0.975335

Obrázek 2.4: Úmrtnostní tabulka (Chakraborty, 2013)

Metoda úmrtnostních tabulek počítá s cenzorovanými událostmi jinak než Kaplanova-Meierova metoda, což můžeme vidět v rovnici (2.12). KM metoda počítá s tím, že v riziku jsou všechny subjekty s cenzorovanou dobou přežití. Na druhou stranu metoda úmrtnostních tabulek předpokládá, že v riziku je průměrně polovina cenzorovaných subjektů.

Jestliže ve stejném čase nastane událost a zároveň cenzura, pak není přesné předpokládat, že všechny události předcházejí všem cenzorováním. Tímto způsobem pracuje Kaplanova-Meierova metoda a oproti metodě úmrtnostních tabulek pravděpodobnost přežití mírně nadhodnocuje. V těchto případech je tedy vhodnější použití metody úmrtnostních tabulek.

Obecně je Kaplanova-Meierova metoda vhodnější pro klinické studie, kde pracujeme s přesnými údaji o době přežití pacientů. Naopak metoda úmrtnostních tabulek je výhodnější v populačních analýzách, kde se úmrtí a cenzorování často vyskytují ve stejném čase (kvůli nepřesnosti zaznamenávání dob přežití či velkému množství pozorovaných subjektů).

Kapitola 3

Semi-parametrické metody

Jak už bylo řečeno, **semi-parametrické metody** jsou jakýmsi přechodem mezi parametrickými a neparametrickými metodami. Pro jejich použití nepotřebujeme konkrétní pravděpodobnostní rozdělení doby přežití T , ale využíváme předpoklady ohledně vysvětlujících parametrů modelu.

3.1 Coxův model

Coxův regresní model neboli model proporcionálních rizik je spolu s Kaplanovým-Meierovým modelem dalším z nejpoužívanějších modelů analýzy přežití. Jeho hlavním pozitivem je jednoduchost použití a následně snadná interpretace výsledků. Tento model se využívá v analýze přežití pro odhadování vlivu dalších (vysvětlujících) proměnných na dobu přežití pacienta.

Kvůli cenzorování používáme modifikované vícerozměrné metody (jako jsou lineární a logistická regrese), sloužící k nalezení závislostí mezi vysvětlovanou a vysvětlujícími proměnnými. Coxova metoda má velice blízko k vícenásobné regresi, kde jako závislou proměnnou bereme rizikovou funkci. Výsledný Coxův regresní model můžeme sumarizovat vyjádřením rizika výskytu sledované události v čase jako funkce vysvětlujících proměnných.

Budeme uvažovat zjednodušený případ, že doby přežití t_i se neopakují, tedy že mezi daty se nevyskytují shody. V praxi je ale běžnější, že se shody v dobách přežití vyskytují. Pak musíme model (respektive metodu odhadu regresních koeficientů) modifikovat (např. Breslow, 1974).

Předpokládejme, že model obsahuje p vysvětlujících proměnných. To znamená, že o každém pacientovi máme k dispozici p údajů. Na počátku máme danou základní rizikovou funkci $h_0(t)$. Naším cílem je charakterizovat daného pacienta, tedy vytvořit model, který by odrážel vliv vybraných vysvětlujících proměnných $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ na tuto základní rizikovou funkci (Collett, 2003). Protože riziková funkce nabývá vždy kladných hodnot, musíme zajistit, že funkce zůstane kladná i po vstupu vysvětlujících proměnných do modelu. Přírozenou pomůckou je využití exponenciální funkce.

Coxův model proporcionálních rizik pak definujeme následovně (Cox, 1972):

$$h(t, X) = h_0(t)e^{\beta'X}, \quad (3.1)$$

kde $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ je vektor regresních koeficientů popisující účinky vysvětlujících proměnných na přežití pacienta (parametrická část modelu) a $h_0(t)$ označuje základní rizikovou funkci, která je stejná pro všechny subjekty a závisí na čase (neparametrická část modelu). Předpokládáme, že tato funkce je libovolná nezáporná funkce a odpovídá rizikové funkci, když $x = 0$.

Regresní koeficienty β_k , $k = 1, 2, \dots, p$ představují vliv vysvětlujících proměnných x_k , $k = 1, 2, \dots, p$ na riziko výskytu sledované události. Přesněji řečeno, β_k představuje číselnou hodnotu, o kterou se navýší přirozený logaritmus rizikové funkce, když se hodnota k -té proměnné zvýší o jednu jednotku za předpokladu, že ostatní vysvětlující proměnné zůstanou stejné. Kladná hodnota regresního koeficientu nám říká, že riziko sledované události je vyšší u pacienta s větší hodnotou odpovídající vysvětlující proměnné. Naopak záporné znaménko regresního koeficientu znamená, že daná vysvětlující proměnná s vyšší hodnotou má protektivní účinek, tedy že riziko výskytu sledované události je menší.

3.1.1 Poměr rizik

Jedním ze způsobů, jak lze interpretovat výstup Coxova modelu, je tzv. **poměr rizik HR** (*hazard ratio*). Pro jakékoliv dva subjekty s vektory proměnných $\mathbf{X}_1 = X_{11}, X_{12}, \dots, X_{1p}$ a $\mathbf{X}_2 = X_{21}, X_{22}, \dots, X_{2p}$ je definován následovně (Klein, 2003)

$$\text{HR} = \frac{h(t, \mathbf{X}_1)}{h(t, \mathbf{X}_2)} = \frac{h_0(t) \exp(\beta' \mathbf{X}_1)}{h_0(t) \exp(\beta' \mathbf{X}_2)} = e^{\beta'(\mathbf{X}_1 - \mathbf{X}_2)}. \quad (3.2)$$

Poměr rizik tedy vyjadřuje, jaký je poměr rizik dvou subjektů s různými proměnnými. Navíc, pokud hodnotíme vliv pouze jediné vysvětlující dichotomické proměnné, která nabývá hodnot jen 0 nebo 1, potom poměr rizik vyjádříme vztahem $\text{HR} = e^{\beta'}$.

Protože poměr rizik nezávisí na základní rizikové funkci (je tedy v čase konstantní), je důležité, aby tato nezávislost byla dodržena i v datech. Graficky to znamená, že rizikové funkce dvou subjektů jsou úměrné a nekříží se, tzn. riziko události jednoho subjektu je násobkem rizika u druhého subjektu po celou dobu pozorování. Tento jev nazýváme **proporcionalita rizik**. Dodržení tohoto předpokladu je zásadní při vytváření Coxova modelu a jeho porušení může vést k nesprávným výsledkům (podhodnocení nebo nadhodnocení poměru rizik).

3.1.2 Odhad regresních koeficientů

Parametry modelu β_i se odhadují nezávisle na základní rizikové funkci $h_0(t)$. Zkoumáním hodnot vektoru β lze zjistit, které vysvětlující proměnné mají vliv na průběh rizikové funkce a křivku přežití prakticky stejně jako v lineární regresní analýze.

Metoda parciální věrohodnosti

Kvůli semi-parametrické podstatě modelu používáme pro odhad regresních koeficientů ne-standardní metody jako je **metoda parciální věrohodnosti** (*partial likelihood method*), kde místo standardní funkce věrohodnosti maximalizujeme **parciální věrohodnostní funkci**, která je závislá jen na vysvětlujících proměnných β .

Budeme uvažovat n pacientů s k pozorovanými dobami přežití. Kvůli očekávanému cenzorování je $k \leq n$. Necht' M_i ($i = 1, 2, \dots, k$) je množina pacientů v riziku v čase t_i pro seřazené pozorované doby přežití $t_1 < t_2 < \dots < t_k$. Pro jednoduchost předpokládáme, že v každém okamžiku t_i nastala událost pouze pro jednoho pacienta, označme jej i a příslušný vektor proměnných délky p tohoto pacienta označme \mathbf{X}_i . Potom $h(t_i, \mathbf{X}_i)$ bude riziková funkce pro pacienta i . Podmíněná pravděpodobnost, že u pacienta i z množiny M_i dojde k výskytu události v čase t_i , za podmínky, že v čase t_i skutečně událost nastala, je definována jako (Klein, 2003):

$$\begin{aligned} P(\text{pacient } i \text{ zemře v čase } t_i | \text{ v čase } t_i \text{ došlo k jedné události}) = \\ = \frac{h(t_i, \mathbf{X}_i)}{\sum_{j \in R_i} h(t_i, \mathbf{X}_j)} = \frac{h_0(t_i) \exp(\beta' \mathbf{X}_i)}{\sum_{j \in R_i} h_0(t_i) \exp(\beta' \mathbf{X}_j)} = \frac{\exp(\beta' \mathbf{X}_i)}{\sum_{j \in R_i} \exp(\beta' \mathbf{X}_j)}. \end{aligned} \quad (3.3)$$

Za předpokladu, že pozorované doby přežití jsou navzájem nezávislé, získáme vynásobením těchto pravděpodobností přes všechny časy přežití parciální věrohodnostní funkci:

$$L(\beta) = \prod_{i=1}^k \frac{\exp(\beta' \mathbf{X}_i)}{\sum_{j \in R_i} \exp(\beta' \mathbf{X}_j)}. \quad (3.4)$$

Pro jednodušší výpočetní operace zavedeme ještě logaritmus parciální věrohodnostní funkce:

$$\ln L(\beta) = l(\beta) = \sum_{i=1}^k \left\{ \beta' \mathbf{X}_i - \ln \left(\sum_{j \in R_i} \exp(\beta' \mathbf{X}_j) \right) \right\} = \sum_{i=1}^k l_i, \quad (3.5)$$

kde l_i značí příspěvek k $l(\beta)$ odpovídající pacientu i .

Maximalizací funkce $l(\beta)$ pak získáme příslušný odhad vektoru regresních koeficientů $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$. To provedeme tak, že položíme příslušné derivace $l(\beta)$ podle β_1 až β_p rovny nule a vyřešíme příslušné soustavy rovnic:

$$\frac{\partial l(\beta)}{\partial \beta_m} = \sum_{i=1}^k \left\{ X_{im} - \frac{\sum_{j \in R_i} X_{jm} \exp(\beta' \mathbf{X}_j)}{\sum_{j \in R_i} \exp(\beta' \mathbf{X}_j)} \right\} = 0. \quad (3.6)$$

Skórový vektor a informační matice

Zavedeme dvě charakteristiky, které souvisí s logaritmem parciální věrohodnosti a používají se pro testování hypotéz o regresních koeficientech $\beta_1, \beta_2, \dots, \beta_p$. Jednou z nich je **skórový vektor** $U_r(\beta)$, který je tvořen složkami (Hosmer, Lemeshow a May, 2008)

$$U_r(\beta) = \sum_{i=1}^k \frac{\partial l_i}{\partial \beta_r}, \quad r = 1, 2, \dots, p, \quad (3.7)$$

kde l_i je logaritmus parciální věrohodnosti v čase t_i .

Druhou charakteristikou je **informační matice** $I(\boldsymbol{\beta})$, která je typu $p \times p$ a na pozici (r, s) obsahuje zápornou hodnotu druhé smíšené parciální derivace $l(\boldsymbol{\beta})$, tedy

$$I_{r,s}(\boldsymbol{\beta}) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s}. \quad (3.8)$$

Navíc platí

$$I(\hat{\boldsymbol{\beta}})^{-1} = \hat{\text{var}}(\hat{\boldsymbol{\beta}}), \quad (3.9)$$

kde $I(\hat{\boldsymbol{\beta}})^{-1}$ je inverzní informační matice a $\hat{\text{var}}(\hat{\boldsymbol{\beta}})$ označuje kovarianční matici maximálně věrohodného odhadu vektoru regresních koeficientů $\boldsymbol{\beta}$ (Hosmer, Lemeshow a May, 2008).

3.1.3 Interval spolehlivosti pro regresní koeficienty

Pro neznámé regresní koeficienty $\beta_1, \beta_2, \dots, \beta_p$ můžeme sestavit $100(1 - \alpha)\%$ interval spolehlivosti s využitím $1 - \frac{\alpha}{2}$ kvantilu standardizovaného normálního rozdělení $z_{1-\frac{\alpha}{2}}$, který bude mít tvar

$$\langle \hat{\beta}_k - z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_k); \hat{\beta}_k + z_{1-\frac{\alpha}{2}} SE(\hat{\beta}_k) \rangle, \quad (3.10)$$

kde SE je standardní chyba odhadu a lze ji získat jako odmocninu prvku z diagonály kovarianční matice $\hat{\text{var}}(\hat{\boldsymbol{\beta}})$, kterou jsme zavedli vztahem (3.9).

Když interval spolehlivosti obsahuje nulu, je to jakýsi indikátor situace, že platnost nulové hypotézy ($H_0 : \beta_k = 0$) nejspíše nezamítneme.

3.1.4 Testy hypotéz pro regresní koeficienty

Chceme rozhodnout, jestli vysvětlující proměnná má statisticky významný vliv na dobu přežití pacienta či nikoliv. Budeme tedy zkoumat hodnotu odpovídajícího regresního koeficientu β_k . Pokud bude tento koeficient nulový, pak přímo ze vztahu (3.2) vyplývá, že poměr rizik, který odpovídá dané proměnné se bude rovnat 1. A to nám říká, že přítomnost této proměnné u pacienta nezvyšuje ani nesnižuje riziko výskytu události.

Nulovou a odpovídající alternativní hypotézu o tom, jestli má regresní koeficient pro k -tou proměnnou statisticky významný vliv na riziko výskytu události či nikoliv, můžeme zapsat následovně:

$$H_0 : \beta_k = 0$$

$$H_1 : \beta_k \neq 0,$$

což lze ekvivalentně zapsat jako nulovou a alternativní hypotézu pro poměr rizik k -té proměnné:

$$H_0 : HR_k = 1$$

$$H_1 : HR_k \neq 1.$$

Pro testování hypotéz o významnosti parametrů β můžeme použít jeden z následujících tří běžně používaných testů.

Test pomocí poměru parciálních věrohodností

Testovací statistika G testu pomocí poměru parciálních věrohodností je dána vztahem (Kleinbaum & Klein, 2005):

$$G = 2 \left(l(\hat{\beta}) - l(0) \right), \quad (3.11)$$

kde $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$ je vektor odhadnutých parametrů, $l(0) = -\sum_{i=1}^k \ln(n_i)$ a n_i je počet subjektů v riziku v čase $t_{(i)}$.

Statistika G má za předpokladu platnosti nulové hypotézy asymptoticky χ^2 rozdělení s jedním stupněm volnosti (obecně lze najednou testovat r regresních koeficientů, pak by testová statistika měla χ^2 rozdělení s r stupni volnosti).

Waldův test

Waldův test souvisí s maximálně věrohodným odhadem první složky odhadnutého vektoru koeficientů $\hat{\beta}$, tedy $\hat{\beta}_k$. Testovací statistika Q_W je dána poměrem odhadu regresního koeficientu a odpovídající standardní chyby, jež získáme jako odmocninu prvku z diagonály kovarianční matice $\hat{var}(\hat{\beta})$, kterou jsme definovali vztahem (3.9). Statistika Q_W je tedy definována vztahem (Hosmer, Lemeshow a May, 2008):

$$Q_W = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \quad (3.12)$$

a řídí se standardizovaným normálním rozdělením (obecně můžeme tento test použít pro testování r koeficientů současně, pak by Q_W měla složitější vyjádření a řídila by se asymptoticky χ^2 rozdělením s r stupni volnosti).

Skóre test

Třetím testem je skóre test, který je založen na podílu derivace přirozeného logaritmu parciální věrohodnosti za platnosti nulové hypotézy ($\beta_k = 0$) a odmocniny prvku na pozici k, k pozorované informační matice (také v bodě $\beta_k = 0$). Testovací statistika Q_S je tedy dána vztahem (Hosmer, Lemeshow a May, 2008)

$$Q_S = \left[\frac{\frac{\partial l(\beta)}{\partial \beta_k}}{\sqrt{-\frac{\partial^2 l(\beta)}{\partial \beta_k^2}}} \right]_{\beta_k=0} \quad (3.13)$$

a má standardizované normální rozdělení. Když toto zobecníme na testování r regresních koeficientů současně bude statistika Q_S mít χ^2 rozdělení s r stupni volnosti, stejně jako u Waldova testu.

3.2 Sestavení Coxova modelu

V předchozí sekci jsme definovali Coxův model, nyní se budeme zabývat tím, jak jej použít v praxi. Ukážeme si, jak vybrat vysvětlující proměnné, které do modelu zahrnout.

Jedním z nejdůležitějších kroků při sestavování modelu je vhodný výběr vysvětlujících proměnných. Vybíráme z těch proměnných, u kterých předpokládáme vliv na přežití pacientů, avšak ve výsledném modelu se pak vyskytují pouze ty s nejsilnějším vlivem. V této souvislosti je zásadním tématem složitost modelu. Vždy by mělo platit, že výsledný model by měl dobře reprezentovat data, ale zároveň by měl být i klinicky dobře interpretovatelný.

Nejpoužívanější metody pro výběr proměnných do Coxova modelu jsou tyto následující procedury nebo jejich kombinace (Pavlík, 2015):

- **metoda postupného přidávání proměnných** (*forward selection*) – začínáme s modelem, který obsahuje jednu nebo pouze málo proměnných, o kterých víme, že mají významný vliv na přežití (např. stádium nádoru). Tuto sadu postupně rozšiřujeme s využitím nějakého kritéria.
- **metoda postupné eliminace proměnných** (*backward elimination*) spočívá v tom, že z úplného modelu, který obsahuje všechny proměnné, jež jsme schopni do modelu zahrnout, postupně odebíráme jednotlivé proměnné. To provádíme na základě nějakého kritéria, například testováním statistické významnosti jednotlivých proměnných.

Pokud máme pro modelování jen malý počet vysvětlujících proměnných a je tedy možné je do modelu zahrnout všechny, nemusíme využít žádné metody pro jejich redukci. Na druhou stranu, jestliže máme vysvětlujících proměnných mnoho, musíme využít vícerozměrné metody pro výběr přijatelného množství proměnných. Cílem použití vícerozměrných metod je odstranit nadbytečné informace, které jsou obsaženy v korelovaných proměnných.

Kapitola 4

Regresní diagnostika

Regresní diagnostika tvoří důležitou součást statistických metod, která posuzuje správnost a vhodnost modelu, tedy zda model správně popisuje vztah mezi souborem proměnných a pozorovaným výsledkem. Pokud použijeme nesprávně navržený model, můžeme dospět k matoucím nebo zcela nesprávným závěrům. Kontrola regresního modelu je tedy stejně důležitým krokem jako samotné sestavení regresního modelu.

Regresní diagnostika se zabývá například zkoumáním splnění předpokladů modelu, hodnocením jeho struktury a hledáním pozorování, která jsou daným modelem špatně vyhodnocena (odlehlá pozorování), nebo která mají poměrně značný vliv na hodnoty regresních koeficientů (vlivné body).

4.1 Analýza reziduí

Rezidua modelu obecně představují rozdíl mezi pozorovanou a předpokládanou hodnotou sledované veličiny. Ve statistice využíváme analýzu reziduí například k testování normality dat, homoskedasticity rozptylu, nezávislosti reziduí (tedy vhodnosti modelu), či výskytu vybočujících hodnot. Při analýze přežití nás bude zajímat analýza reziduí především z hlediska testování vhodnosti modelu a pak také pro hledání vlivných a odlehlých pozorování.

Vysoké hodnoty reziduí znamenají, že model špatně vysvětluje cílovou proměnnou pomocí proměnných vysvětlujících. Také jejich „systematické chování“, ač pouze v nízkých absolutních hodnotách, je známkou špatného modelu, především pak nesprávného výběru vysvětlujících proměnných. Většinou totiž není vhodné do lineárního modelu zařadit všechny proměnné, lepší může být použití kvadratické formy modelu nebo danou proměnnou kategorizovat. Kategorizaci spojitých proměnných bychom měli provádět obezřetně, jelikož jde o ztrátu informace, která může způsobit zvýšení variability modelu a vést ke zkreslení výsledků. Často je však tato úprava nutná kvůli interpretaci závěrů (například kategorizace věku na: 0–50 let, 51–60 let, 61–70 let, 71 a více let).

Rezidua modelu přežití by se měla chovat náhodně, protože reprezentují formu nevysvětlené variability. Obecně, rezidua odpovídající statistickým chybám by měla vykazovat přibližně symetrické rozložení okolo nulové střední hodnoty a rozptyl by měl být kon-

stantní shora ohraničený. Pokud se tedy podíváme na grafy reziduí, neměl by být patrný žádný trend, tzn. rezidua by měla vytvářet rovnoměrný vodorovný pás.

Odchyly od výše zmíněných předpokladů většinou znamenají fakt, že rezidua obsahují nějakou informaci, jež nám říká, že model je sestavený špatně.

V analýze přežití je interpretace hodnot reziduí komplikovanější kvůli přítomnosti cenzorování. Pomoci může grafické znázornění a vyhlazení trendu (např. jádrové), kvůli asymetričnosti reziduí. Následující rezidua jsou při analýze přežití nejpoužívanější. Jejich využití je obecnější, proto místo dosud užívaného označení „pacient“ zavedeme tato rezidua pro širší pojem – „subjekt“.

4.1.1 Martingale rezidua

Martingale reziduum u sledovaného subjektu definujeme jako rozdíl mezi pozorovaným a očekávaným počtem událostí. Očekávaný počet událostí \hat{E}_i je vyjádřen kumulativní rizikovou funkcí do doby t_i . Pro i -tý subjekt je tedy reziduum dáno vztahem (Hosmer, Lemeshow a May, 2008):

$$M_i(t) = d_i - \hat{E}_i = d_i - \hat{H}_0(t_i) \exp(\mathbf{X}_i' \hat{\boldsymbol{\beta}}), \quad (4.1)$$

kde d_i je počet událostí subjektu i v čase t (v datech analýzy přežití tedy může nabývat hodnot 1 nebo 0).

Hodnoty martingale reziduí jsou kladné pro subjekty, jejichž doba přežití byla kratší než jsme predikovali. Naopak záporné hodnoty odpovídají subjektům s delší skutečnou dobou přežití. Tato rezidua můžeme použít i ke grafickému hodnocení toho, zda je model vhodný. Pokud mají rezidua nulovou střední hodnotu, pak je model vhodný.

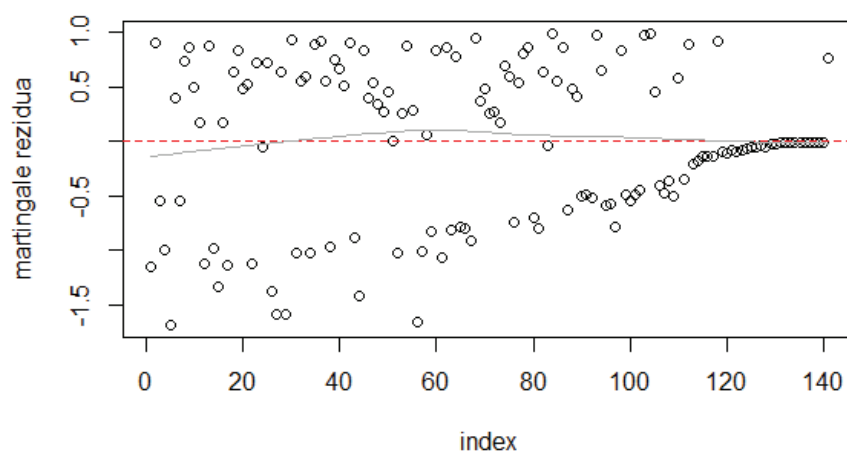
Martingale rezidua můžeme využít také pro identifikaci nelineárního vlivu námi vybrané vysvětlující proměnné a pro vyhodnocení, jestli např. daná proměnná nebyla z modelu vyloučena nesprávně. Pro názornou ukázkou nalezneme tato rezidua na Obrázku 4.1 na straně 23. Jedná se o rezidua jednorozměrného Coxova modelu s proměnnou GTV, která byla vypočtena z našich reálných dat. Tento model bychom vzhledem k martingale reziduům označili za vhodný.

4.1.2 Deviační rezidua

Martingale rezidua mají jednu nevýhodu – nejsou symetrická a maximální hodnota, které dosahují je jedna. Tento nedostatek můžeme odstranit transformací, pak získáme tzv. deviační rezidua.

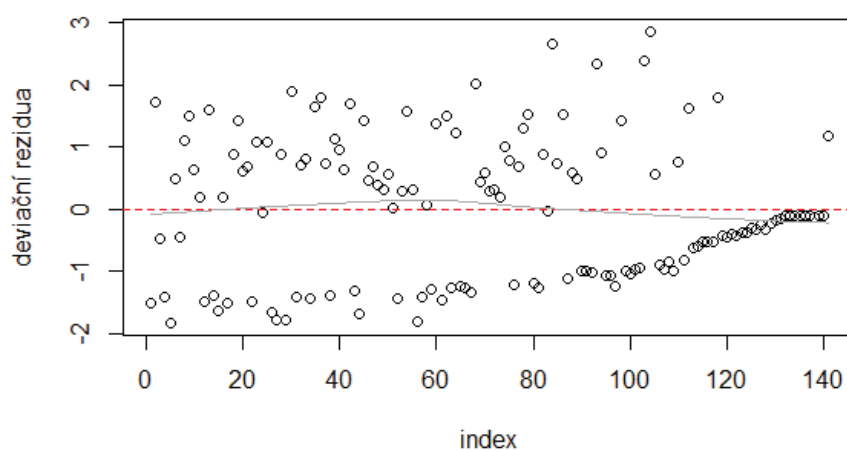
Deviační rezidua mají nulovou střední hodnotu, stejně jako martingale rezidua. Liší se však v tom, že jsou symetricky rozložena okolo nuly a tedy i snadněji interpretovatelná. Pro i -tý subjekt jsou definována následovně (Marubini a Valsecchi, 1995)

$$D_i(t) = \text{sgn}(M_i(t)) \sqrt{-2 [M_i(t) + d_i(t) \ln(d_i(t) - M_i(t))]}, \quad (4.2)$$



Obrázek 4.1: Martingale rezidua pro proměnnou „GTV (ml) - objem primárního nádoru“

kde $M_i(t)$ představuje martingale reziduum pro i -tý subjekt pacienta v čase t a $d_i(t)$ je počet pozorovaných událostí pro subjekt i .



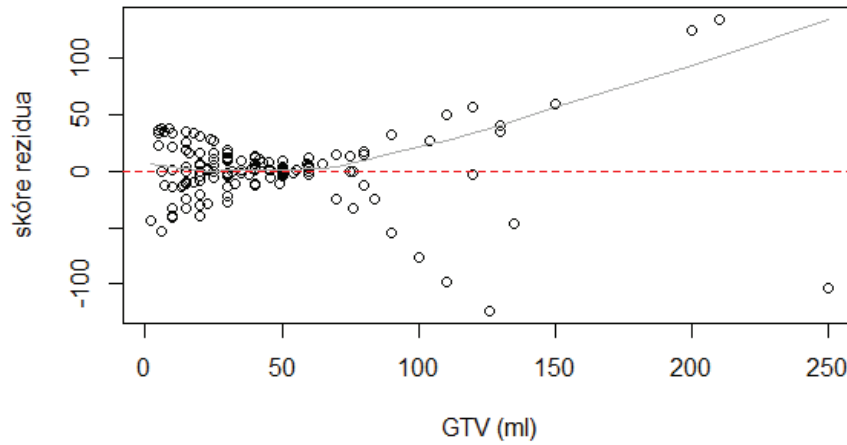
Obrázek 4.2: Deviační rezidua pro proměnnou „GTV (ml) - objem primárního nádoru“

Tato rezidua se používají především k identifikaci odlehlých pozorování, tedy těch subjektů, u kterých je značný rozpor mezi pozorovaným a předpokládaným rizikem. To si vysvětlíme v podkapitole 4.3 na straně 29. Na Obrázku 4.2 vidíme, jak může vypadat graf těchto reziduů. Takto vykreslená rezidua predikují vhodný model.

Martingale i deviační rezidua představují jednu číselnou hodnotu na jeden studovaný subjekt.

4.1.3 Skóre rezidua

Skóre rezidua na rozdíl od výše zmíněných dvou typů představují číselný vektor obsahující pro každý hodnocený subjekt jedno číslo pro každou proměnnou zahrnutou do modelu. Tato rezidua udávají odhad změny regresních koeficientů v případě, že by daný subjekt byl z modelu vyřazen. Součet všech skóre reziduů je roven nule. Vysoká hodnota skóre rezidua indikuje „vlivné pozorování“, tj. subjekt, který má velký příspěvek k odhadu regresních koeficientů.



Obrázek 4.3: Skóre rezidua pro proměnnou „GTV (ml) - objem primárního nádoru“

Pro výpočet použijeme martingale rezidua a první derivaci logaritmu parciální věrohodnosti definovaného v předchozí kapitole vztahem (3.5). Pro m -tou proměnnou i -tého subjektu jsou rezidua dána rovnicí (Hosmer, Lemeshow a May, 2008)

$$L_{im} = \sum_{j=1}^k (X_{im} - \bar{X}_{wjm}) dM_i(t_j), \quad (4.3)$$

kde X_{im} představuje hodnotu m -té proměnné pro i -tý subjekt, \bar{X}_{wjm} je její vážený průměr v množině subjektů M_i a $dM_i(t_j)$ je definována jako:

$$dM_i(t_j) = dd_i(t_j) - Y_i(t_j) \exp(\mathbf{X}'_j \boldsymbol{\beta}) h_0(t_j) \quad (4.4)$$

a značí změnu v martingale reziduu pro i -tý subjekt v čase t_j . Číslo $dd_i(t_j)$ je změna v tzv. čítací funkci pro i -tý subjekt v čase t_j . Pro cenzorované doby se vždy rovná nule a pro necenzorované nabývá hodnoty nula ve všech časech kromě toho, ve kterém byla

pozorována událost, tzn. $dd_i(t_i) = 1$. Funkce $Y_i(t_j)$ je indikátor, zda je subjekt v riziku, tzn. je rovna jedné až do doby, kdy nastane událost, tedy pro $t_j \leq t_i$ je $Y_i(t_j) = 1$. A hodnota $h_0(t_j)$ představuje základní riziko v čase t_j . Situaci si můžeme prohlédnout na Obrázku 4.3. Skóre rezidua vykreslujeme oproti hodnotám dané proměnné. Na obrázku je patrný trend, který je způsoben malým počtem subjektů pro vysoké hodnoty GTV.

4.1.4 Schoenfeldova rezidua

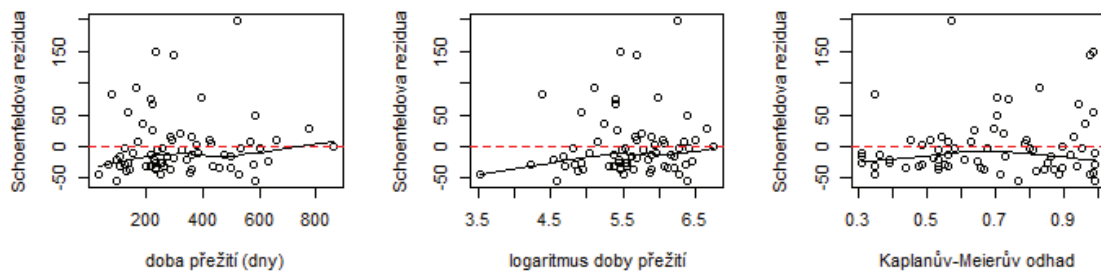
Schoenfeldova rezidua jsou stejně jako skóre rezidua tvořena vektorem čísel, tzn. každé proměnné náleží pro jeden subjekt jedna hodnota.

Tato rezidua se vztahují k pozorované době přežití a v této době představují rozdíl mezi pozorovanou hodnotou proměnné X_k u subjektu s nastalou událostí a očekávanou hodnotou této proměnné, kterou můžeme vypočítat z dat všech subjektů v riziku v daném čase t . Součet všech Schoenfeldových reziduí je roven nule.

Schoenfeldova rezidua definujeme pomocí příspěvku subjektů k derivaci logaritmu parciální věrohodnosti. Pro m -tou proměnnou i -tého subjektu jsou dána následovně (Lee & Wang, 2003):

$$R_{im} = \delta_i(X_{im} - \bar{X}_{w_i m}) = \delta_i \left(X_{im} - \frac{\sum_{j \in R_i} X_{jm} \exp(\mathbf{X}'_j \boldsymbol{\beta})}{\sum_{j \in R_i} \exp(\mathbf{X}'_j \boldsymbol{\beta})} \right), \quad (4.5)$$

kde X_{im} představuje hodnotu m -té proměnné pro i -tý subjekt, $\bar{X}_{w_i m}$ je vážený průměr proměnné v množině R_i . Číslo δ je indikátor cenzury, což zde znamená, že Schoenfeldova rezidua definujeme pouze pro necenzorovaná pozorování.



Obrázek 4.4: Schoenfeldova rezidua pro proměnnou „GTV (ml) - objem primárního nádoru“

I v tomto případě by rezidua měla být bez trendu náhodně rozmístěna ve vodorovném pásu kolem nuly. Příklad těchto reziduí znázorňuje Obrázek 4.4. Tato rezidua vykreslujeme oproti hodnotám času, logaritmu času nebo proti Kaplanově-Meierově odhadu.

V transformované formě, kdy pro škálování použijeme odhad rozptylu Schoenfeldových reziduí, je můžeme použít k testování předpokladu proporcionality rizik. Více o škálovaných reziduích můžeme nalézt např. v publikaci od Hosmera, Lemeshowa a May, 2008.

4.2 Hodnocení vhodnosti modelu

Hodnocení vhodnosti modelu je důležitý a velice obtížný úkol především kvůli tomu, že obecně není specifikováno, co vhodný model je a co není, tzn. zda vhodně popisuje soubor dat, či nikoliv.

4.2.1 Grønnesbyho a Borganův test

Pro hodnocení celkové vhodnosti modelu můžeme použít testy dobré shody, například Grønnesbyho-Borganův test (May a Hosmer, 2004). Ten spočívá v tom, že množství pozorovaných událostí v předem dané skupině má být zhruba stejné jako množství událostí očekávaných. Pak je model považován za správný.

Grønnesbyho-Borganův test využívá výše zavedená martingale rezidua. Nejdříve se data uspořádají a roztrídí do skupin podle hodnot předpokládaného rizikového skóre. Pro každou skupinu se vypočítá suma reziduí. Jestliže jsou splněny všechny předpoklady a model je maximálně vhodný pro datový soubor, součet reziduí pro každou skupinu bude roven nule. Následně se testuje nulová hypotéza předpokládaných nulových sum reziduí v každé skupině.

Grønnesbyho-Borganův test můžeme realizovat také prostřednictvím skórového testu. Grønnesbyho-Borganův test je algebraicky ekvivalentní odvozenému testu, jež se opírá o testování významnosti přidané indikátorové proměnné pomocí skórového testu. Subjekty se opět roztrídí do skupin na základě jejich rizikového skóre, ze kterých se jedna skupina vybere za referenční. V dalším kroku se pro každý subjekt v každé skupině (kromě referenční skupiny) vytvoří hodnoty pro indikátorové proměnné a testuje se významnost jejich přidání prostřednictvím skórového testu. Testuje se hypotéza, jestli se hodnoty koeficientů příslušící indikátorovým proměnným rovnají nule. Jestliže je daný model pro data vhodný, nemělo by jej přidání indikátorových proměnných významně ovlivnit (koeficienty jsou nulové).

Grønnesbyho-Borganův test má velice blízko k testu vhodnosti modelu pro logistickou regresi (Hosmer a Lemeshow, 1999), kde jsou pozorování rozdělena do deseti skupin podle předpokládaných pravděpodobností. Poté se porovnávají pozorované a předpokládané hodnoty výsledné proměnné v každé skupině na základě chí-kvadrát statistiky s $k - 2$ stupni volnosti.

Ověřuje se hypotéza, že rozdíly mezi pozorovanými a předpokládanými hodnotami výsledné proměnné jsou nulové v každé skupině. Kvůli podobnosti mezi testy bylo navrženo roztrídění subjektů do deseti skupin. Případně, aby počet předpokládaných událostí v každé

skupině nebyl menší než pět (základní předpoklad pro použití chí-kvadrát statistiky), pak by počet skupin neměl být větší, než počet událostí vydělený pěti.

4.2.2 Akaikeho informační kritérium

Akaikeho informační kritérium slouží k porovnání více modelů tím, že posuzuje jejich schopnost vysvětlit variabilitu v pozorovaných datech (Akaike, 1974). Pro každý z porovnávaných modelů počítáme hodnotu kritéria

$$AIC = -2l(\boldsymbol{\beta}, (t_1, d_1), (t_2, d_2), \dots, (t_n, d_n)) + 2(c + a), \quad (4.6)$$

kde $l(\boldsymbol{\beta}, (t_1, d_1), (t_2, d_2), \dots, (t_n, d_n))$ je logaritmus věrohodnostní funkce modelu, c značí počet vysvětlujících proměnných v modelu a a je počet parametrů daného pravděpodobnostního rozdělení.

Akaikeho informační kritérium tedy porovnává věrohodnost a zároveň složitost modelu. Čím nižší hodnotu AIC model má, tím je vhodnější. Obecně platí, že čím více proměnných je zahrnuto, tím roste věrohodnost modelu ($l(\boldsymbol{\beta}, (t_1, d_1), (t_2, d_2), \dots, (t_n, d_n))$) a tím se hodnota kritéria AIC snižuje. Na druhou stranu, čím větší počet proměnných, tím vyšší bude hodnota druhé části vzorce ($2(c + a)$) a tím víc bude růst i hodnota AIC . Celkem to znamená, že Akaikeho informační kritérium znázorňuje rovnováhu mezi věrohodností a složitostí modelu.

Nevýhodou AIC je však to, že pro Coxův model nelze použít. Je to kvůli tomu, že Coxův model je metoda semi-parametrická, tudíž nespecifikuje konkrétní pravděpodobnostní rozdělení dob přežití. Srovnání Coxova modelu by bylo velmi zavádějící i proto, že tento model využívá metodu parciální věrohodnosti.

4.3 Verifikace modelu

Důležitým předpokladem pro korektnost použití Coxova modelu je provedení následujících bodů:

- testování dodržení proporcionality rizik v čase (HR je konstantní po celou dobu trvání studie),
- vyhodnocení vlivných pozorování a jejich celkového dopadu na parametry modelu.

4.3.1 Hodnocení proporcionality rizik

Klíčovým předpokladem Coxova modelu je proporcionalita rizik každých dvou pozorovaných subjektů. Proporcionalita rizik říká, že efekt daných proměnných (neboli rizikových faktorů) se v čase nemění. Je tedy velmi důležité ověřit, jestli tento předpoklad dané proměnné splňují. Pokud nikoliv, není vhodné použít Coxův model pro daná data a je potřeba hledat jiné možnosti analýzy. Druhou možností je nalézt postup, jak data upravit, aby byl

předpoklad splněn. Porušením proporcionality rizik je odhad poměru rizik HR pro tento faktor nadhodnocen nebo podhodnocen.

Ověření předpokladu proporcionality rizik můžeme vyhodnocovat dvěma způsoby, a to numericky nebo graficky. K dispozici máme mnoho metod. Existují metody, které vyžadují rozdělení dob přežití nebo kategorizaci vysvětlujících proměnných, jiné lze použít přímo na nezměněný původní datový soubor. Je čistě jen na uživateli, kterou metodu si vybere.

Grafický způsob ověření proporcionality rizik je nejjednodušší. Zde budeme požadovat, aby se Kaplanovy-Meierovy křivky přežití rovnoměrně vzdalovaly a nikdy se nepřekřížily. Existují ale i propracovanější metody ověření proporcionality rizik:

- **Grafické ověření pomocí kumulativní rizikové funkce**

Použití logaritmu kumulativní rizikové funkce je vhodnější než porovnávání křivek přežití (Pavlík, 2015). Kumulativní riziková funkce poskytuje jednoduchou pomůcku pro ověření jednorozměrné proporcionality rizik, protože funkce přežití vzhledem k proměnné k v případě Coxova modelu musí splňovat následující rovnici

$$S_k(t) = \exp(-H_0(t) \exp(x_k \beta_k)), \quad (4.7)$$

a po zlogaritmování tedy platí

$$\ln(-\ln S_k(t)) = \ln H_0(t) + x_k \beta_k. \quad (4.8)$$

Jestliže bude předpoklad proporcionality rizik splněn, křivky pro jednotlivé hodnoty proměnné budou přibližně rovnoběžné. Vizuální ověřování ale nezohledňuje více faktorů zároveň, proto tuto metodu nemůžeme použít pro hodnocení proporcionality rizik u modelů s více než dvěma proměnnými.

- **Test pomocí časově závislé proměnné (*time-dependent covariate test*)**

Když přidáme do modelu časově závislou proměnnou, kterou získáme transformací původní proměnné (značíme $x_k(t) = x_k g(t)$), dovolí nám to, aby se poměr rizik měnil v čase (Ng'andu, 1997). Potom testujeme, zda regresní koeficient takto přidané proměnné je statisticky významný.

Jestliže koeficient, který přísluší této časově závislé proměnné, bude statisticky významně různý od nuly, pak předpoklad proporcionality rizik neplatí. Výsledek testu nám tedy říká, že se v modelu vyskytuje časová závislost poměru rizik příslušného k -té proměnné.

- **Test založený na škálovaných Schoenfeldových reziduiích (*test based on scaled Schoenfeld residuals*)**

Jestliže je v případě k -té proměnné předpoklad proporcionality rizik splněn, měla by být rezidua při znázornění proti času náhodně rozdělena okolo nuly (Marubini, Valsecchi, 1995). A to tak, že stejně v časech na počátku, uprostřed i na konci sledování.

Pokud tedy proložíme regresní křivku škálovanými Schoenfeldovými rezidui a budeme sledovat časový trend (popř. nějaké schodovité změny či zlomy), pak můžeme proporcionalitu rizik zamítnout.

- **Test pomocí Pearsonova korelačního koeficientu** (*test based on Pearson correlation coefficient*)

Početní ověření proporcionality rizik můžeme provést pomocí Pearsonova korelačního koeficientu mezi časy přežití a škálovanými Schoenfeldovými rezidui. Obvykle je testová statistika používaná pro testování korelačního koeficientu založena na t-testu, nicméně, pro analýzu přežití používá většina softwarů chí-kvadrát test (odvozeno např. v Grambsch, Therneau, 1994).

4.3.2 Identifikace vlivných pozorování

Dalším měřítkem, zda je model správný, je vyhodnocení vlivných pozorování. V analýze přežití používáme obdobné metody jako v lineární a logistické regresi (Hosmer, Lemeshow a May, 2008).

V klasickém modelu identifikujeme vlivná pozorování (tj. subjekty se špatně predikovanými hodnotami) na základě tzv. pákových bodů (*leverage values*). Leverage je diagnostická statistika, která měří rozdíly hodnot proměnných jednotlivých subjektů a proměnných celého souboru a vypočítá se jako vzdálenost mezi hodnotou jedné proměnné a celkovým průměrem proměnných. V analýze přežití se používá aproximace leverage statistiky, a to skóre rezidua (Kapitola 4.1, strana 24). Jako vlivná pozorování označíme ty body, které jsou v grafu skóre reziduí vzdálené od ostatních (Hosmer, Lemeshow a May, 2008). Tato pozorování pak musíme zkoumat podrobněji.

V lineární a logistické regresi vysoká hodnota leverage statistiky sama o sobě ještě nemusí nic znamenat. Nicméně, subjekty s vysokou hodnotou pro leverage mohou příliš ovlivňovat odhad regresních parametrů.

Pro vyšetření vlivných pozorování můžeme použít další metodu, a to Cookovu vzdálenost. Ta počítá, jak se mění hodnoty odhadnutých koeficientů po odstranění konkrétního subjektu z dat. Pokud je hodnota Cookovy vzdálenosti pro konkrétní subjekt nízká, pak to znamená, že právě tento subjekt odhad parametrů modelu výrazně neovlivňuje. V analýze přežití se používá aproximace Cookovy vzdálenosti, a to škálovaná skórová rezidua (viz např. Hosmer, Lemeshow a May, 2008).

4.3.3 Identifikace odlehlých hodnot

Subjekty, u kterých pozorujeme výskyt události příliš brzy nebo velmi pozdě vzhledem k ostatním subjektům s podobnými vstupními hodnotami proměnných, nazýváme odlehlá pozorování. K jejich identifikaci můžeme použít dříve zavedená deviační rezidua (viz Kapitola 4.1, strana 22), pomocí kterých graficky identifikujeme odlehlá pozorování.

Kapitola 5

Analýza reálných dat

V této kapitole budeme aplikovat dříve popsané metody na konkrétní onkologická data, která poskytla Fakultní nemocnice Ostrava (FNO). Informace o pacientech byly zaznamenávány od května roku 2008 do června 2013. Máme k dispozici údaje o 141 pacientech zaznamenávající dobu do úmrtí. U každého pacienta bylo sledováno mnoho proměnných. V následujícím seznamu uvádíme ty proměnné, které MUDr. Jakub Cvek, Ph.D. vybral jako vhodné k vyhodnocení.

- pohlaví: žena/muž,
- věk,
- prognóza dle NCCN (National Comprehensive Cancer Network), vysvětleno níže,
- Dg dle MKN10 (Mezinárodní klasifikace nemocí a přidružených zdravotních problémů, desátá revize):
 - C01 = Zhoubný novotvar (ZN) kořene jazyka
 - C03 = ZN dásně
 - C04 = ZN ústní spodiny
 - C08 = ZN jiných a neurčených slinných žláz
 - C09 = ZN mandle - tonzily
 - C11 = ZN nosohltanu (nazofaryngu)
 - C13 = ZN hypofaryngu
 - C14 = ZN jiných a nepřesně určených lokalizací rtu, ústní dutiny a hltanu
 - C31 = ZN vedlejších dutin
 - C32 = ZN hrtanu
- GTV (ml) – objem primárního nádoru,
- GTVLU (ml) – objem uzlinové metastázy,
- GTV+GTVLU (ml),

- KI – výkonnostní stav: 50 až 100 (viz Tabulka 5.1),
- grading – stupeň diferenciacie: 0 = nezjištěno, 1 = dobře diferencovaný karcinom, 2 = středně diferencovaný karcinom, nebo 3 = málo diferencovaný karcinom,
- fyzikální dávka radioterapie,
- biologicky efektivní dávka,
- doba radioterapie: dny,
- dávka na frakci: 1,40, 1,45 nebo 1,50

V Tabulce 5.2 na další straně je souhrnný výpis těchto parametrů a převedení ordinálních dat na kategoriální, pomocí rozdělení hodnot do intervalů. Proměnnou *Dg dle MKN10* jsme po konzultaci s lékařem rozdělili do čtyř skupin podle diagnóz, které spolu věcně souvisí.

hodnota	popis
100	normální stav, bez obtíží, bez známek onemocnění
90	schopen normální aktivity, velmi málo symptomů
80	schopen normální aktivity s mírnými obtížemi, symptomy
70	soběstačný, neschopen normálních aktivit či práce
60	občas vyžaduje pomoc, zvládá většinu osobních potřeb
50	často potřebuje pomoc, často vyžaduje lékařskou léčbu
40	invalidita, vyžaduje speciální péči a pomoc
30	těžká invalidita, indikace k přijetí do nemocnice
20	těžce nemocný, potřeba akutní hospitalizace, podpůrných opatření a léčby
10	moribundní, rychle progredující fatální onemocnění
0	smrt

Tabulka 5.1: Karnofského index (KI), (Karnofsky & kol., 1948). Stupnice je popsána v intervalech po 10, lékař však může podle potřeby měnit skóre i v rozmezí těchto hodnot.

Kategorie důkazů a konsensů (*NCCN Categories of Evidence and Consensus*) rozdělených podle *NCCN Guidelines® & Clinical Resources*:

1. Na základě důkazů vysoké úrovně panuje podle NCCN jednotný konsensus (shoda), že zákrok je vhodný.
2. Na základě důkazů nízké úrovně panuje podle NCCN konsensus (shoda), že zákrok je vhodný.
3. Na základě důkazů jakékoli úrovně panuje podle NCCN neshoda, zda je zákrok vhodný či nikoliv.

název proměnné	varianty	počet úmrtí	celkem pacientů
pohlaví	žena	15	34
	muž	55	107
věk	[40, 56]	14	29
	(56, 59]	13	23
	(59, 67]	21	49
	(67, 90]	22	40
prognóza dle NCCN	1	43	89
	2	17	32
	3	10	20
Dg dle MKN10	C01+C03+C04	24	37
	C08+C11+C31	2	7
	C09	15	40
	C13+C14+C32	29	57
GTV (ml)	[0, 50]	47	104
	(50, 100]	13	24
	(100, 150]	7	10
	(150, 250]	3	3
GTVLU (ml)	[0, 5]	39	83
	(5, 20]	20	34
	(20, 50]	5	13
	(50, 100]	2	4
	(100, 200]	4	7
GTV+GTVLU (ml)	[0, 50]	33	79
	(50, 100]	21	35
	(100, 150]	10	18
	(150, 200]	3	6
	(200, 250]	3	3
KI - výkonnostní stav	[50, 70]	12	21
	(70, 89]	38	77
	(89, 100]	20	43
Grading	0	23	49
	1	13	25
	2	27	50
	3	7	17
fyzikální dávka radioterapie	(62, 72]	13	39
	(72, 74)	39	65
	[74, 77)	18	37
biologicky efekt. dávka	[64, 72)	17	35
	[72, 75)	34	66
	[75, 80)	19	40
doba radioterapie (dny)	[30, 35)	5	16
	[35, 40)	53	109
	[40, 45]	12	16
dávka na frakci	1,40	12	38
	1,45	40	71
	1,50	18	32

Tabulka 5.2: Popis proměnných

5.1 Neparametrická analýza

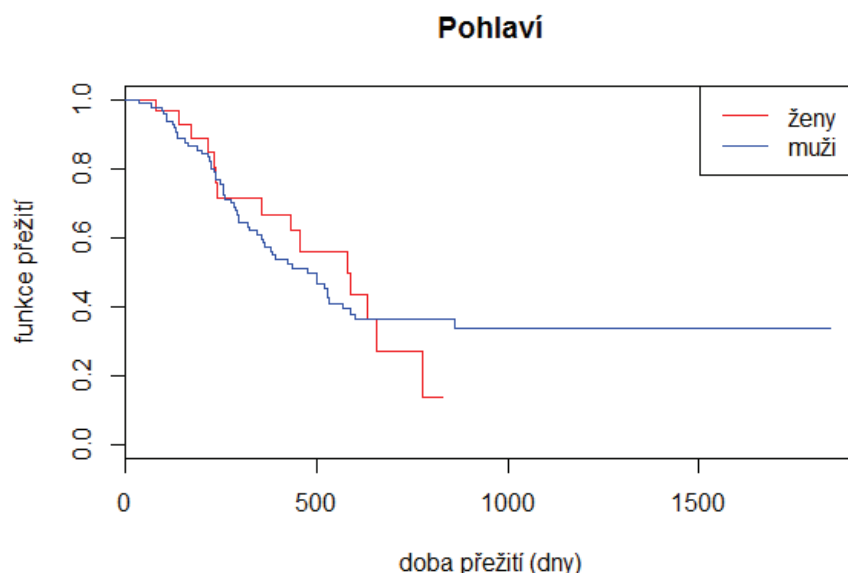
Nejprve budeme neparametrickými metodami porovnávat vliv různých proměnných na dobu přežití. Kaplanův-Meierův odhad funkce přežití nám ukáže grafické srovnání jednotlivých variant proměnných. Testování rozdílnosti časů přežití v jednotlivých skupinách provedeme pomocí log-rank testu – funkce `survdiff` z balíčku `survival`. Uvedeme také mediány dob přežití a jejich intervaly spolehlivosti (IS), které spočítáme pomocí funkce `survfit` ze stejného balíčku.

5.1.1 Log-rank test

Budeme testovat nulovou hypotézu H_0 o shodnosti dvou či více křivek přežití proti alternativní hypotéze H_1 , která tvrdí, že mezi křivkami se vyskytuje statisticky významný rozdíl (Kapitola 2, strana 12). Dále spočítáme medián doby přežití, tedy dobu, které se dožije polovina sledovaných pacientů (Kapitola 2, strana 12). Pokud medián, resp. hranice intervalu spolehlivosti mediánu nebudou uvedeny, znamená to, že se příslušný odhad funkce přežití, resp. spolehlivostní pásmo vyskytuje v celém definičním oboru nad hranicí 0,5 odhadované pravděpodobnosti přežití.

Vliv pohlaví

Na Obrázku 5.1 vidíme srovnání Kaplanových-Meierových křivek přežití pro muže a ženy. Do výzkumu bylo zařazeno 34 žen a 107 mužů. Nulová hypotéza pro log-rank test říká, že



Obrázek 5.1: Kaplan-Meierův odhad funkce přežití zvlášť pro ženy a pro muže

funkce přežití pro obě pohlaví jsou stejné. P-hodnota nám vyšla 0,981, tedy podstatně větší než hladina významnosti 0,05, tzn. nulovou hypotézu o shodnosti obou křivek nezamítáme.

Zjistili jsme, že pohlaví nemá statisticky významný vliv na dobu přežití. Tento výsledek jsme očekávali, neboť KM křivky se výrazně neliší.

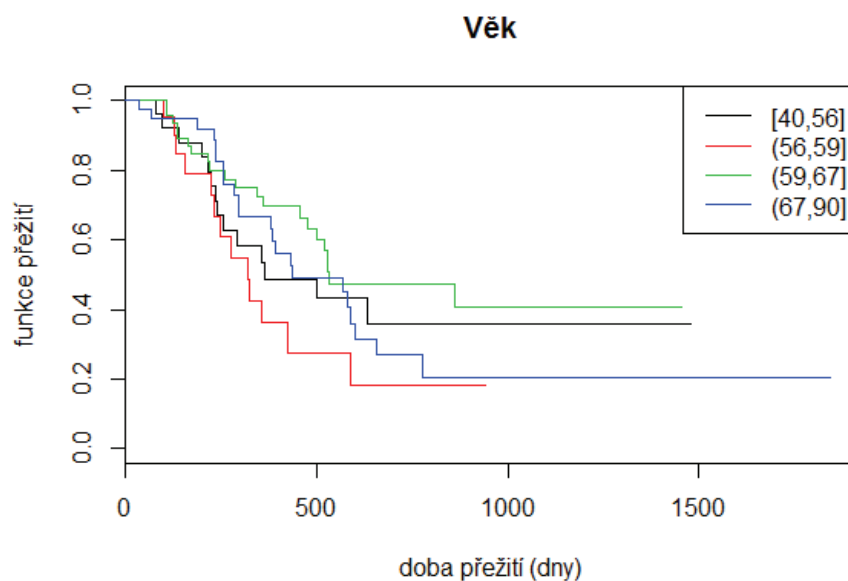
V Tabulce 5.3 uvádíme mediány dob přežití spolu s 95% intervaly spolehlivosti. Hodnoty mediánů jsou relativně blízké, takže se potvrdilo, že pohlaví neovlivňuje dobu přežití.

varianta	medián doby přežití (dny)	95% IS
žena	582	(431, –)
muž	476	(359, 601)

Tabulka 5.3: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou pohlaví

Vliv věku

Na Obrázku 5.2 máme srovnání Kaplanových-Meierových křivek přežití pro čtyři věkové skupiny. Četnosti a počet úmrtí v jednotlivých skupinách nalezneme v Tabulce 5.2 v úvodu této kapitoly. Log-rank test testuje nulovou hypotézu, že křivky pro všechny věkové sku-



Obrázek 5.2: Kaplan-Meierův odhad funkce přežití zvlášť pro čtyři věkové skupiny

piny jsou shodné. P-hodnota nám vyšla 0,135, tedy vyšší než hladina významnosti 0,05, tzn. test říká, že věk nemá statisticky významný vliv na dobu přežití.

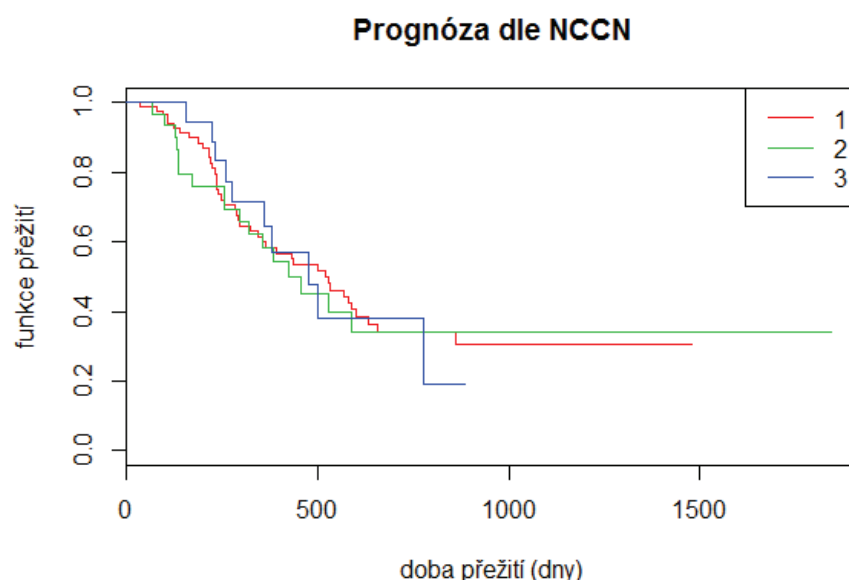
V Tabulce 5.4 jsou uvedeny mediány dob přežití a jejich 95% intervaly spolehlivosti.

varianta	medián doby přežití (dny)	95% IS
[40, 56]	363	(256, –)
(56, 59]	319	(230, –)
(59, 67]	531	(476, –)
(67, 90]	436	(379, 777)

Tabulka 5.4: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou věk

Vliv prognózy dle NCCN

Obrázek 5.3 znázorňuje srovnání Kaplanových-Meierových křivek přežití pro tři skupiny pacientů podle prognóz dle NCCN. Log-rank opět testuje nulovou hypotézu o shodnosti



Obrázek 5.3: Kaplan-Meierův odhad funkce přežití pro skupiny pacientů podle prognózy dle NCCN

křivek přežití. P-hodnota nám vyšla velice vysoká: 0,978, tedy nulovou hypotézu nezamítáme. Výsledek testu tvrdí, že nezáleží, zda má pacient prognózu podle NCCN kategorie 1, 2 nebo 3 (interpretace jednotlivých kategorií nalezneme na straně 31). Tento výsledek jsem mohli očekávat, protože všechny tři křivky mají velice podobný průběh.

V Tabulce 5.5 jsou uvedeny mediány dob přežití a jejich 95% intervaly spolehlivosti. Mediány dob přežití leží velice blízko sebe, tedy odpovídají výsledkům testování log-rank testem.

varianta	medián doby přežití (dny)	95% IS
1	521	(357, 657)
2	458	(319, –)
3	476	(359, –)

Tabulka 5.5: Mediány doby přežití a 95% IS mediánů pro proměnnou prognóza dle NCCN

Vliv diagnózy dle MKN10

V našich datech se vyskytuje deset různých diagnóz (C01, C03, C04, C08, C09, C11, C13, C14, C31, C32). Vzhledem k celkovému počtu 141 pacientů je však těchto skupin příliš mnoho a je nutné agregování. Po konzultaci s lékařem jsme vytvořili 4 skupiny pacientů s diagnózami, které spolu věcně souvisí (C01+C03+C04, C08+C11+C31, C09, C13+C14+C32). Tabulka 5.6 uvádí příslušné mediány dob přežití a jejich 95% intervaly spolehlivosti.

varianta	medián doby přežití (dny)	95% IS
C01, C03, C04	426	(237, 587)
C08, C11, C31	357	(109, –)
C09	657	(379, –)
C13, C14, C32	528	(363, –)

Tabulka 5.6: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou diagnóza dle MKN10

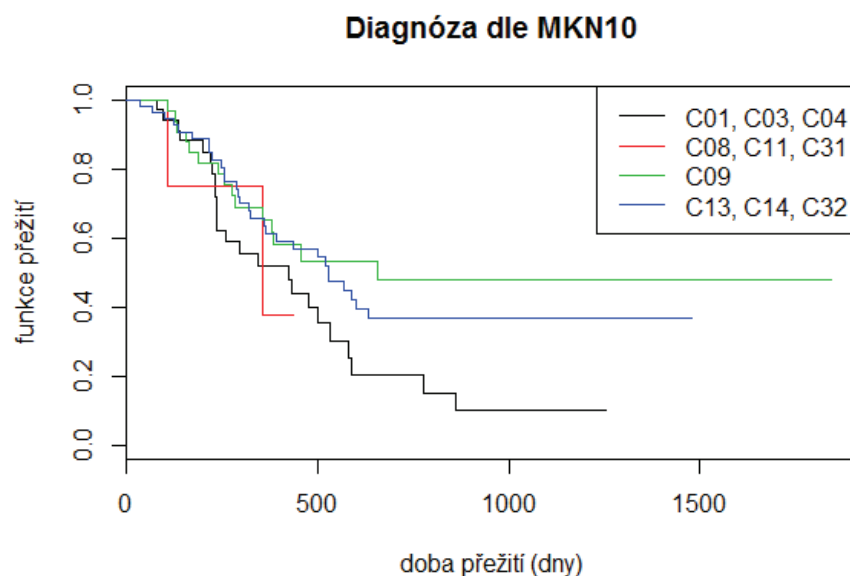
Na Obrázku 5.4 jsou znázorněny Kaplanovy-Meierovy křivky přežití pro tyto 4 skupiny. Log-rank testem budeme opět porovnávat křivky přežití. P-hodnota nám vyšla 0,123, tedy větší než hladina významnosti 0,05, proto nulovou hypotézu nezamítáme. Log-rank testem jsme došli k závěru, že vliv diagnózy dle MKN10 na dobu přežití pacientů není statisticky významný.

Vliv GTV – objemu primárního nádoru

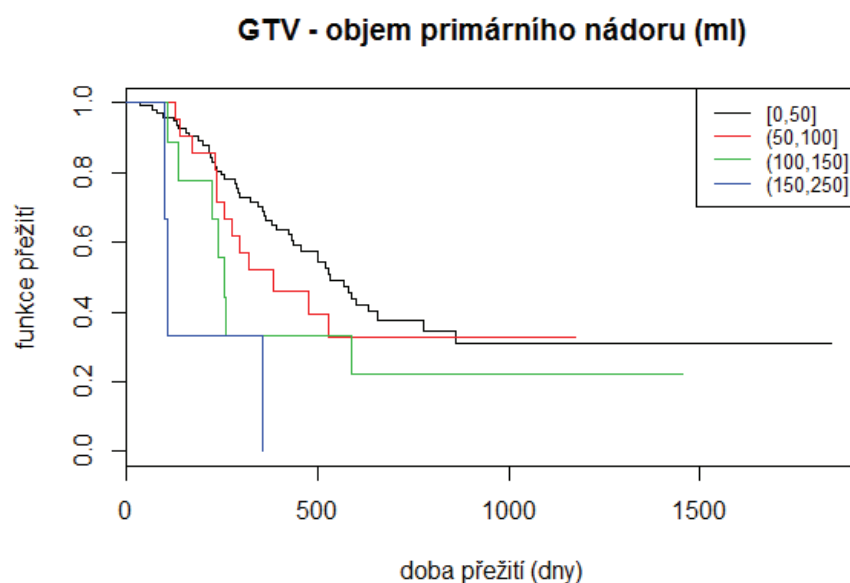
Obrázek 5.5 znázorňuje porovnání Kaplanových-Meierových křivek přežití pro čtyři skupiny pacientů podle objemu primárního nádoru. V Tabulce 5.7 nalezneme příslušné mediány dob přežití spolu s intervaly spolehlivosti. Můžeme si povšimnout, že mediány se od sebe podstatně liší.

varianta	medián doby přežití (dny)	95% IS
[0, 50]	531	(436, 777)
(50, 100]	384	(254, –)
(100, 150]	256	(222, –)
(150, 250]	109	(98, –)

Tabulka 5.7: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou GTV



Obrázek 5.4: Kaplan-Meierův odhad funkce přežití zvlášť pro různé skupiny diagnóz



Obrázek 5.5: Kaplan-Meierův odhad funkce přežití podle GTV

Log-rank testem testujeme nulovou hypotézu o shodnosti všech čtyř křivek přežití. P-hodnota nám vyšla 0,005, tedy nulovou hypotézu zamítáme. Zjistili jsme, že se křivky liší a že záleží na tom, jak velký je objem primárního nádoru. Dále otestujeme, které dvě konkrétní varianty tuto odlišnost způsobují. Použijeme vícenásobné porovnávání, tedy log-rank test pro jednotlivé dvojice. Výsledky si můžeme prohlédnout v Tabulce 5.8. Vidíme, že statisticky významný rozdíl se prokázal u dvojic křivek pro objemy primárního

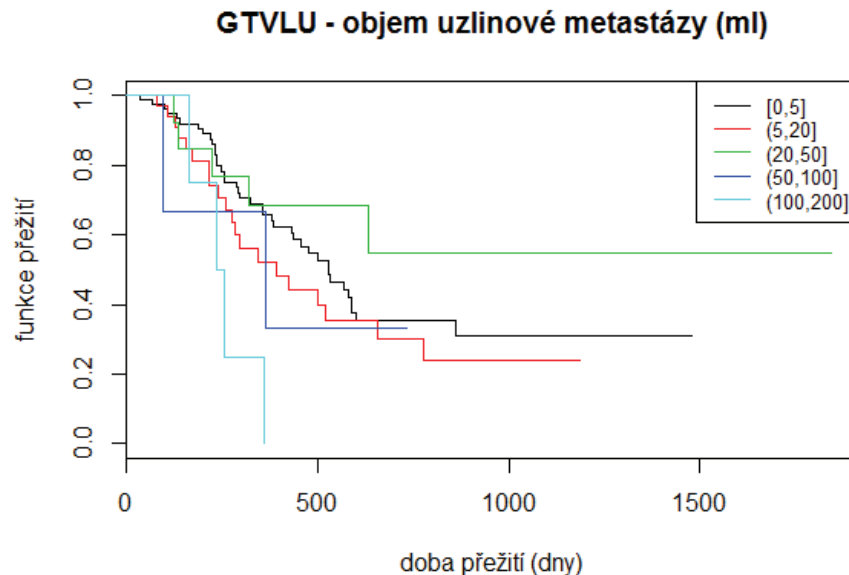
dvojice intervalů (ml)	hodnota testovací statistiky	p-hodnota
[0, 50], (50, 100]	0,5	0,476
[0, 50], (100, 150]	2,0	0,160
[0, 50], (150, 250]	12,8	<0,001
(50, 100], (100, 150]	0,7	0,390
(50, 100], (150, 250]	5,2	0,022
(100, 150], (150, 250]	1,5	0,223

Tabulka 5.8: Log-rank test pro dvojice intervalů objemů primárního nádoru (ml)

nádoru $[0, 50]$ – černá (nejnižší hodnoty) a $(150, 250]$ – modrá (nejvyšší hodnoty) a také pro dvojici intervalů $(50, 100]$ – červená (druhý nejnižší interval hodnot) a $(150, 250]$ – modrá (nejvyšší hodnoty). Tyto výsledky jsou patrné i z grafu. Můžeme tvrdit, že doba přežití pacientů s nejvyššími hodnotami GTV, $(150, 250]$, je statisticky významně kratší než doba přežití pacientů s nízkými hodnotami GTV, konkrétně z intervalu $[0, 100]$.

Vliv GTVLU – objemu uzlinové metastázy

Na Obrázku 5.6 jsou znázorněny Kaplanovy-Meierovy křivky přežití pro pět skupin pacientů rozdělených podle objemů uzlinové metastázy. V Tabulce 5.9 jsou mediány dob přežití spolu s jejich intervaly spolehlivosti. Log-rank testem opět testujeme nulovou



Obrázek 5.6: Kaplan-Meierův odhad funkce přežití zvlášť pro různé hodnoty GTVLU (ml)

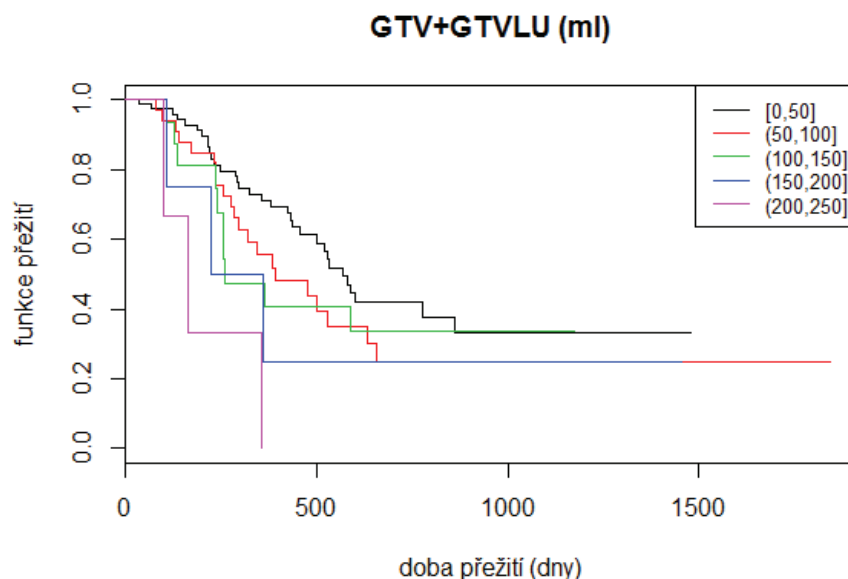
hypotézu o shodnosti křivek přežití. P-hodnota nám vyšla 0,078, tedy větší než hladina významnosti 0,05, proto nulovou hypotézu nezamítáme. Podle tohoto testu nemá objem uzlinové metastázy statisticky významný vliv na dobu přežití.

varianta	medián doby přežití (dny)	95% IS
[0, 5]	529	(431, 860)
(5, 20]	394	(276, –)
(20, 50]	-	(319, –)
(50, 100]	363	(94, –)
(100, 200]	246	(163, –)

Tabulka 5.9: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou GTVLU

Vliv GTV + GTVLU

Na Obrázku 5.7 jsou znázorněny Kaplanovy-Meierovy křivky přežití pro pět skupin pacientů v závislosti na součtu objemů primárního nádoru (GTV) a uzlinové metastázy (GTVLU). Jde o lineární kombinaci předchozích dvou proměnných, nicméně lékaři tuto hodnotu zaznamenávají zvlášť, proto ji zde pro úplnost uvedeme také. V Tabulce 5.10 jsou uvedeny mediány dob přežití a jejich spolehlivostní intervaly.



Obrázek 5.7: Kaplan-Meierův odhad funkce přežití pro součet hodnot GTV a GTVLU (ml)

Log-rank testem budeme opět testovat nulovou hypotézu o shodnosti křivek přežití. Dostali jsme p-hodnotu rovnu 0,021, tedy menší než hladina významnosti 0,05, proto nulovou hypotézu zamítáme. To znamená, že součet objemů primárního nádoru a uzlinové metastázy statisticky významně ovlivňuje dobu přežití pacientů. Dále budeme zjišťovat, které konkrétní křivky se odlišují.

V Tabulce 5.11 vidíme, že odlišnosti mezi křivkami na hladině významnosti 0,05

varianta	medián doby přežití (dny)	95% IS
[0, 50]	570	(458, –)
(50, 100]	394	(297, –)
(100, 150]	261	(240, –)
(150, 200]	290	(109, –)
(200, 250]	163	(98, –)

Tabulka 5.10: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou GTV+GTVLU

dvojice intervalů (ml)	hodnota testovací statistiky	p-hodnota
[0, 50], (50, 100]	1,9	0,168
[0, 50], (100, 150]	0,7	0,394
[0, 50], (150, 200]	1,0	0,308
[0, 50], (200, 250]	12,8	<0,001
(50, 100], (100, 150]	0,0	0,985
(50, 100], (150, 200]	0,3	0,596
(50, 100], (200, 250]	4,9	0,027
(100, 150], (150, 200]	0,3	0,604
(100, 150], (200, 250]	2,9	0,091
(150, 200], (200, 250]	1,4	0,235

Tabulka 5.11: Log-rank test pro dvojice intervalů proměnné GTV+GTVLU (ml)

se prokázaly v případě pro intervaly [0, 50] – černá (nejnižší hodnoty) a (200, 250] – růžová (nejvyšší hodnoty) a pro (50, 100] – červená (druhý nejnižší interval hodnot), (200, 250] – růžová (nejvyšší hodnoty). Můžeme říct, že doba přežití pacientů s nejvyššími hodnotami GTV+GTVLU, (200, 250], je statisticky významně kratší než doba přežití pacientů s nízkými hodnotami GTV+GTVLU, konkrétně z intervalu [0, 100].

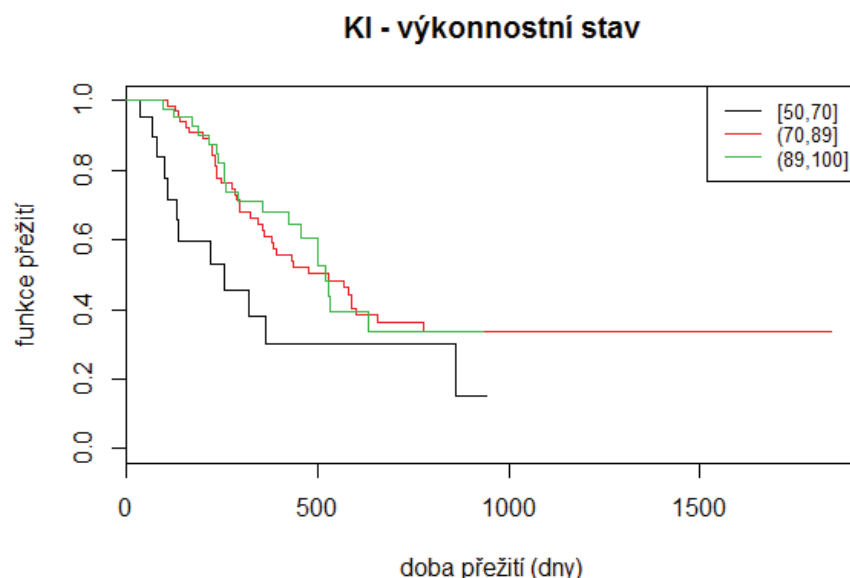
Vliv KI – výkonnostního stavu

Na Obrázku 5.8 jsou znázorněny tři Kaplanovy-Meierovy křivky přežití pro pacienty, které jsme rozdělili do skupin podle jejich hodnoty KI – výkonnostního stavu. V Tabulce 5.12 najdeme hodnoty mediánů dob přežití spolu s jejich intervaly spolehlivosti. Vidíme, že medián doby přežití pacientů s hodnotou KI z intervalu [50, 70] je podstatně nižší než ostatní dva mediány. Log-rank testem opět testujeme nulovou hypotézu o shodnosti

varianta	medián doby přežití (dny)	95% IS
[50, 70]	254	(131, –)
(70, 89]	528	(379, 777)
(89, 100]	521	(458, –)

Tabulka 5.12: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou KI

křivek přežití. P-hodnota nám vyšla 0,046, tedy menší než hladina významnosti 0,05,



Obrázek 5.8: Kaplan-Meierův odhad funkce přežití podle výkonnostního stavu

proto tvrdíme, že hodnota výkonnostního stavu ovlivňuje dobu přežití. V Tabulce 5.13 si můžeme prohlédnout výsledky vícenásobného porovnávání jednotlivých intervalů.

dvojice intervalů	hodnota testovací statistiky	p-hodnota
$[50, 70], (70, 89]$	5,0	0,026
$[50, 70], (89, 100]$	4,6	0,032
$(70, 89], (89, 100]$	0,0	0,854

Tabulka 5.13: Log-rank test pro dvojice intervalů výkonnostních stavů

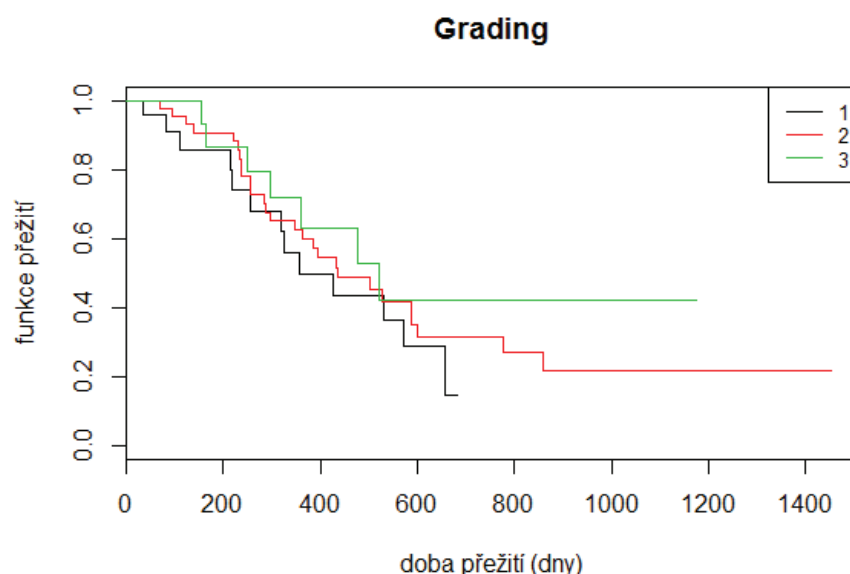
Vidíme, že statisticky významný rozdíl se prokázal u dvojic křivek pro výkonnostní stavy z intervalů $[50, 70]$ a $(70, 89]$ a také pro dvojici intervalů $[50, 70]$ a $(89, 100]$. Černá křivka pro KI z intervalu $[50, 70]$ se statisticky významně liší od červené ($KI \in (70, 89]$) i zelené ($KI \in (89, 100]$) křivky. To znamená, že pacienti s hodnotami KI nižšími než 70 bodů mají významně kratší dobu přežití než pacienti s výkonnostním stavem vyšším než 70. Konkrétní význam hodnot skóre KI je uveden v Tabulce 5.1 na straně 31.

Vliv proměnné grading – stupeň diferenciacce

Na Obrázku 5.9 jsou znázorněny Kaplanovy-Meierovy křivky přežití pro tři skupiny pacientů (1 = dobře diferencovaný karcinom, 2 = středně diferencovaný karcinom nebo 3 = málo diferencovaný karcinom). V původních datech jsou skupiny čtyři, avšak skupina s hodnotou 0 zahrnuje pacienty, u kterých tato hodnota nebyla zjištěna, proto ji do testu nezařazujeme. V Tabulce 5.14 jsou uvedeny mediány dob přežití spolu s 95% spolehlivostními intervaly.

varianta	medián doby přežití (dny)	95% IS
1	357	(255, –)
2	436	(346, 777)
3	521	(359, –)

Tabulka 5.14: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou grading



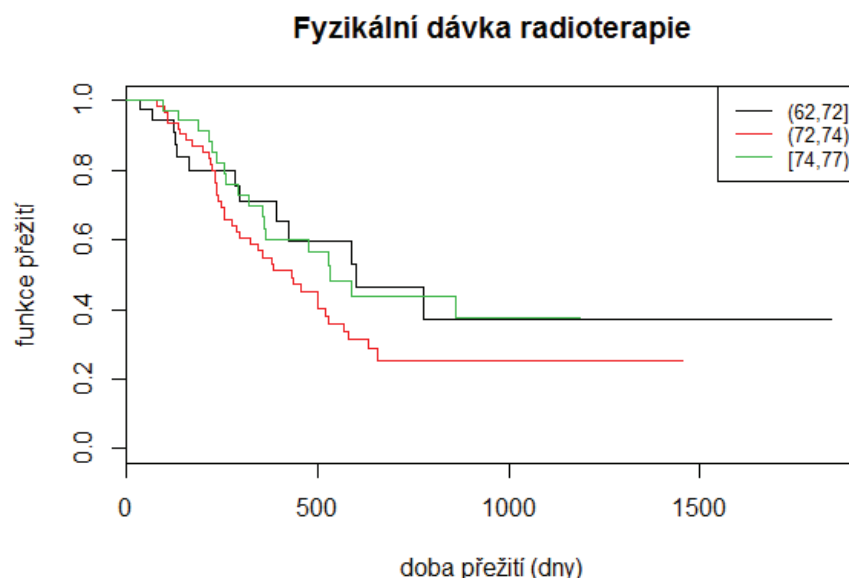
Obrázek 5.9: Kaplan-Meierův odhad funkce přežití zvlášť pro 3 stupně diferenciacie

Pomocí log-rank testu jsme zjišťovali, zda jsou křivky přežití shodné. Podle p-hodnoty 0,40 můžeme říci, že křivky se statisticky významně neliší. Podle grafického znázornění bychom očekávali výsledek jiný, neboť zelená křivka (pro hodnotu 3) leží téměř celá nad černou křivkou (pro hodnotu 1). To by znamenalo, že pacienti s málo diferencovaným karcinomem mají delší dobu přežití než pacienti s dobře diferencovaným karcinomem. Nicméně, log-rank test tento rozdíl vyhodnotil jako statisticky nevýznamný.

Vliv fyzikální dávky radioterapie

Na Obrázku 5.10 jsou znázorněny Kaplanovy-Meierovy křivky přežití pro tři skupiny pacientů podle velikosti fyzikální dávky radioterapie. Tabulka 5.15 uvádí hodnoty mediánů doby přežití a jejich intervaly spolehlivosti.

P-hodnota 0,271 získaná log-rank testem nám udává, že křivky se statisticky významně neliší, tedy doba přežití nezávisí na fyzikální dávce radioterapie. To lze vyčíst i z Obrázku 5.10, kde všechny tři křivky si jsou podobné a vzájemně se kříží.



Obrázek 5.10: Kaplan-Meierův odhad funkce přežití zvlášť pro různé fyzikální dávky radioterapie

varianta	medián doby přežití (dny)	95%IS
(62, 72]	601	(394, –)
(72, 74)	431	(295, 570)
[74, 77)	531	(359, –)

Tabulka 5.15: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou fyzikální dávka radioterapie

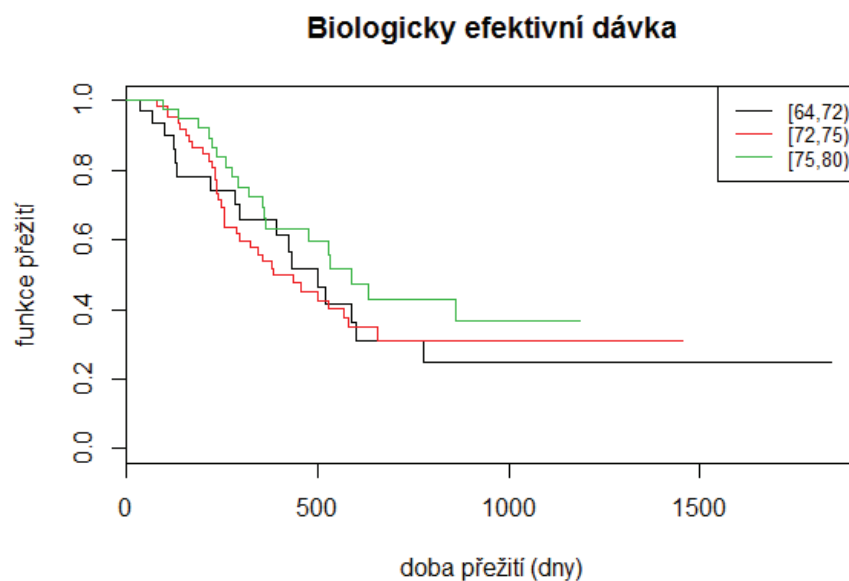
Vliv proměnné „biologicky efektivní dávka“

Na Obrázku 5.11 vidíme srovnání Kaplanových-Meierových křivek přežití pro různá množství biologicky efektivní dávky radioterapie. V Tabulce 5.16 jsou uvedeny mediány dob přežití spolu s intervaly spolehlivosti. Nulová hypotéza pro log-rank test říká, že funkce

varianta	medián doby přežití (dny)	95% IS
[64, 72)	501	(297, –)
[72, 75)	384	(295, 657)
[75, 80)	587	(363, –)

Tabulka 5.16: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou biologicky efektivní dávka

přežití jsou pro všechny tři skupiny stejné. P-hodnota nám vyšla 0,383, tedy větší než hladina významnosti 0,05, tzn. nulovou hypotézu o shodnosti křivek nezamítáme. Výsledek testu tvrdí, že biologicky efektivní dávka nemá statisticky významný vliv na dobu přežití pacientů. Toto je zřejmé i z grafu, kde všechny tři křivky přežití mají velice podobný



Obrázek 5.11: Kaplan-Meierův odhad funkcí přežití podle množství biologicky efektivní dávky

průběh.

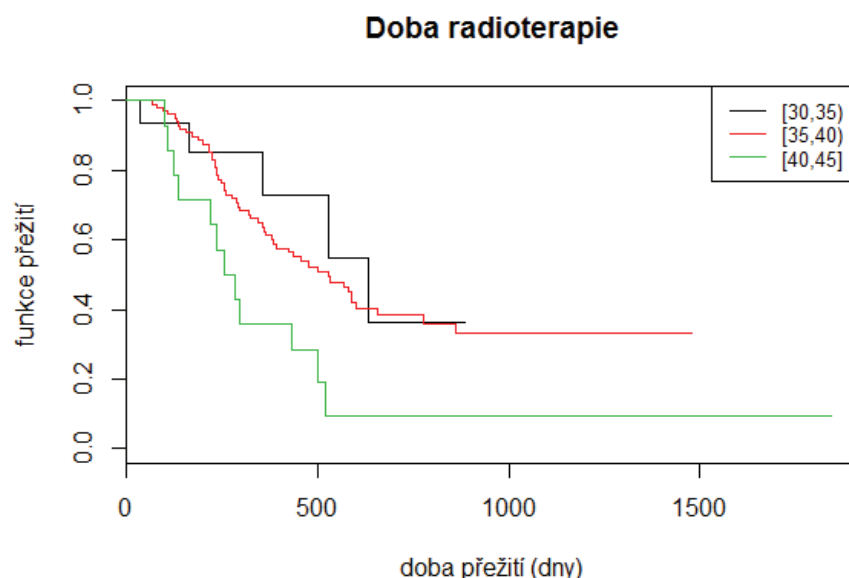
Vliv doby radioterapie

Na Obrázku 5.12 jsou znázorněny tři Kaplanovy-Meierovy křivky přežití pro pacienty, které jsme rozdělili do skupin podle délky doby radioterapie. Tabulka 5.17 uvádí hodnoty mediánů dob přežití a jejich intervaly přežití. Můžeme si povšimnout, že medián doby přežití pacientů s délkou doby radioterapie 40 až 45 dní je podstatně nižší než u ostatních dvou skupin.

varianta	medián doby přežití (dny)	95% IS
[30, 35)	631	(528, –)
[35, 40)	529	(384, 777)
[40, 45]	270	(219, –)

Tabulka 5.17: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou doba radioterapie

Log-rank testem opět testujeme nulovou hypotézu o shodnosti křivek přežití. P-hodnota nám vyšla 0,017, tedy menší než hladina významnosti 0,05, proto nulovou hypotézu o shodnosti křivek zamítáme. V Tabulce 5.18 si můžeme prohlédnout výsledky vícenásobného porovnávání jednotlivých dvojic intervalů.



Obrázek 5.12: Kaplan-Meierův odhad funkce přežití podle doby radioterapie

dvojice intervalů	hodnota testovací statistiky	p-hodnota
[30, 35), [35, 40)	0,2	0,641
[30, 35), [40, 45]	5,1	0,024
[35, 40), [40, 45]	7,2	0,007

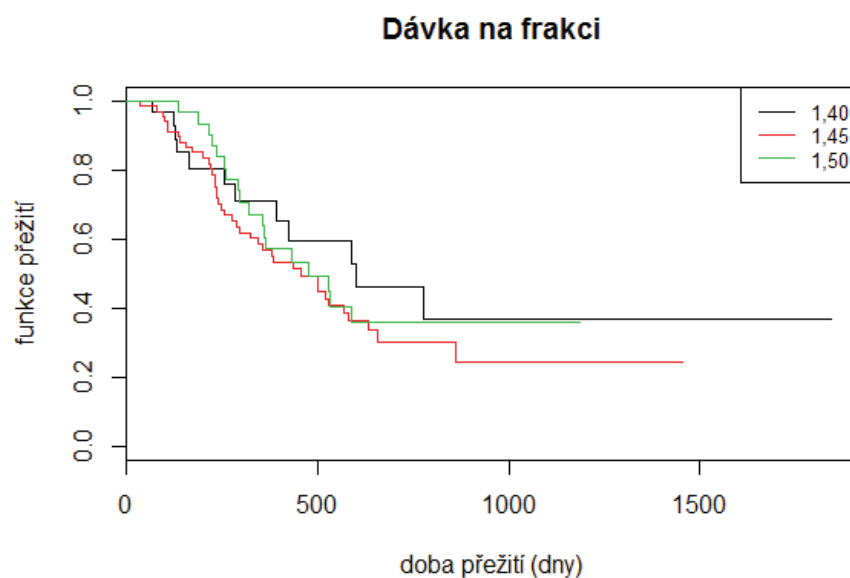
Tabulka 5.18: Log-rank test pro dvojice podle délky doby radioterapie

Vidíme, že statisticky významný rozdíl se prokázal u dvojic křivek pro doby z intervalů $[30, 35)$ a $[40, 45]$ a také pro dvojici intervalů $[35, 40)$ a $[40, 45]$. Tedy třetí skupina (zeleně) se liší od první (černě) i od druhé (červeně). To znamená, že doba přežití pacientů s delší dobou radioterapie (40 až 45 dní) je statisticky významně kratší než pro pacienty, u kterých doba radioterapie trvá méně než 40 dní. Tato skutečnost ovšem může být způsobena také nerovnoměrným množstvím pacientů ve skupinách. V první skupině je 16, ve druhé 109 a ve třetí 16 pacientů.

Vliv dávky na frakci

Na Obrázku 5.13 jsou znázorněny Kaplanovy-Meierovy křivky přežití pro pacienty rozdělené do tří skupin podle velikosti dávky na frakci. V Tabulce 5.19 jsou mediány dob přežití spolu s jejich intervaly spolehlivosti.

Pomocí log-rank testu jsme opět zjišťovali, zda jsou křivky přežití shodné. Podle p-hodnoty 0,533, můžeme říci, že křivky se významně neliší. To znamená, že doba přežití nezávisí na tom, zda dávka na frakci je 1,40, 1,45 nebo 1,50.



Obrázek 5.13: Kaplan-Meierův odhad funkce přežití zvlášť pro 3 různé dávky na frakci

varianta	medián doby přežití (dny)	95% IS
1,40	601	(394, –)
1,45	458	(324, 631)
1,50	476	(358, –)

Tabulka 5.19: Mediány doby přežití a 95% intervaly spolehlivosti mediánů pro proměnnou dávka na frakci

5.2 Semi-parametrická analýza

Nyní budeme pomocí semi-parametrických metod, které byly zavedeny v Kapitole 3, zkoumat vliv všech uvažovaných proměnných současně. Také se budeme zabývat interakcemi mezi nimi. K tomu budeme využívat funkce `coxph` z balíčku `survival`.

5.2.1 Coxův model

Krok 1

Prvním krokem při sestavování Coxova modelu je smysluplný výběr proměnných, které budou vstupovat do prvního vícerozměrného modelu. Tyto proměnné vybíráme na základě jednorozměrných testů. Pro určení statisticky významných proměnných se používá test parciálním věrohodnostním poměrem (TPVP), který jsme zavedli v Kapitole 3 na straně 19. V případě kategoriálních proměnných je log-rank test ekvivalentní s tímto testem.

V tabulce 5.20 je shrnutí výsledků testu parciálním věrohodnostním poměrem pro kategoriální proměnné provedeným pomocí softwaru R a také výsledky log-rank testu z předchozí podkapitoly. Ty proměnné, které budou statisticky významné alespoň na 20% významnostní hladině poté zařadíme do prvního vícerozměrného modelu (Hosmer, Lemeshow a May, 2008).

proměnná	statistika G	df	p-hodnota TPVP	p-hodnota log-rank testu
pohlaví	0,00	1	0,982	0,981
prognóza dle NCCN	0,04	2	0,978	0,978
Dg dle MKN10	5,45	3	0,142	0,121
grading	2,81	3	0,422	0,405
dávka na frakci	1,27	2	0,531	0,534

Tabulka 5.20: Testy pro kategoriální proměnné

V Tabulce 5.21 jsou uvedeny výsledky testu parciálním věrohodnostním poměrem a Waldova testu (Kapitola 3, strana 19) pro spojité proměnné. Z obou testů jsme získali stejné závěry.

proměnná	statistika G	df	p-hodnota TPVP	p-hodnota Waldova testu
věk	0,24	1	0,622	0,624
GTV	4,41	1	0,036	0,022
GTVLU	2,89	1	0,089	0,059
GTV+GTVLU	6,91	1	0,009	0,005
KI	3,86	1	0,049	0,046
fyz.dávka	0,02	1	0,887	0,887
biolog.efekt.dávka	1,56	1	0,211	0,209
doba radioterapie	5,05	1	0,025	0,019

Tabulka 5.21: Testy pro spojité proměnné

Z první Tabulky 5.20 vyplývá, že jediná statisticky významná kategoriální proměnná je *Dg dle MKN10*, z druhé Tabulky 5.21 spojitých proměnných vidíme, že statisticky významné jsou *GTV*, *GTVLU*, *GTV+GTVLU*, *KI* a *doba radioterapie*. Proměnná *biologicky efektivní dávka* překračuje mez 20% jen nepatrně a protože se z praktického hlediska dá předpokládat, že tato proměnná by vliv na dobu přežití mít mohla, zahrneme ji do sestavování modelu. Protože proměnná *GTV+GTVLU* je součet dvou proměnných, které jsou i samostatně významné, nebudeme ji do modelu zařazovat.

Dále si musíme dát pozor na to, že je rozdíl pracovat s kategoriální nebo spojitou proměnnou. Pokud se jedná o kategoriální proměnnou dichotomickou, můžeme s ní pracovat stejně jako se spojitou. Pokud však kategoriální proměnná nabývá více hodnot, musíme skupiny překódovat na nuly a jedničky. Tedy například proměnná *Dg dle MKN10* nebude vstupovat jako jedna proměnná, nýbrž jako více tzv. *dummy* proměnných, označíme: *dg1*, *dg2*, *dg3*, *dg4*. Proměnná *dg1* bude nabývat hodnoty 1, pokud se bude jednat o pacienta s diagnózou z první skupiny (viz Tabulka 5.2, tedy C01, C03 nebo C04), jinak bude obsahovat nuly. Proměnná *dg2* bude mít hodnoty 1 pro pacienty s diagnózami z druhé skupiny (tedy C08, C11 nebo C31), jinak 0. Obdobně pro proměnné *dg3* a *dg4*. Všechny tyto proměnné však do modelu nezahrneme, způsobilo by to perfektní multikolinearitu ($dg1 = 1 - dg2 - dg3 - dg4$). Počet *dummy* proměnných je tedy vždy o 1 nižší než počet původních kategorií. Vynechaná proměnná je tzv. referenční proměnná (zde *dg1*).

Krok 2

Nyní již můžeme sestavit první vícerozměrný model (označíme M1), a to z těchto proměnných: *Dg dle MKN10*, *GTV*, *GTVLU*, *KI*, *biolog. efekt. dávka* a *doba radioterapie*. V Tabulce 5.22 jsou uvedeny odhadnuté koeficienty modelu, z-souřadnice a p-hodnoty Waldova testu. Hodnota $l(M1)$ představuje logaritmus parciální věrohodnostní funkce modelu M1.

proměnná	koeficient $\hat{\beta}_i$	z	p-hodnota
dg2	-0,181	-0,212	0,832
dg3	-0,952	-2,761	0,006
dg4	-0,395	-1,325	0,185
GTV	0,006	1,768	0,077
GTVLU	0,009	2,142	0,032
KI	-0,042	-2,326	0,020
biolog. efekt. dávka	-0,031	-0,647	0,518
doba radioterapie	0,074	1,298	0,194

Tabulka 5.22: Model M1, $l(M1) = -285,513$

Krok 3

V následujícím kroku budeme model redukovat. Jako první odstraníme proměnnou s nejvyšší p-hodnotou nad hladinou významnosti 0,05, tedy proměnnou *biolog. efekt. dávka*. Ve skutečnosti má nejvyšší p-hodnotu varianta *dg2* proměnné *Dg dle MKN10*. Jednu variantu kategoriální proměnné však samostatně vyřadit nemůžeme a protože ostatní dvě

varianty mají p-hodnoty nižší, v modelu je ponecháme.

Takto získaný model nazveme M2 (Tabulka 5.24) a budeme posuzovat, zda odebrání proměnné mělo vliv na změnu modelu. Tuto skutečnost budeme posuzovat pomocí změn hodnot koeficientů (Tabulka 5.23) a testem parciálním věrohodnostním poměrem

$$G = 2(l(M1) - l(M2)) = 2(-285,513 - (-285,719)) = 0,412.$$

Tato statistika má p-hodnotu = 0,520, tedy větší než je hladina významnosti 0,05. A také relativní změny v koeficientech nejsou větší než 20%. Můžeme tedy říci, že odebrání proměnné *biologicky efektivní dávka* nemělo významný vliv na změnu modelu.

proměnná	změna koeficientu (%)
dg2	-1,9
dg3	-1,1
dg4	3,4
GTV	-1,7
GTVLU	-2,9
KI	-2,6
doba radioterapie	17,8

Tabulka 5.23: Změny koeficientů

proměnná	koeficient $\hat{\beta}_i$	z	p-hodnota
dg2	-0,177	-0,208	0,835
dg3	-0,942	-2,741	0,006
dg4	-0,408	-1,375	0,169
GTV	0,006	1,732	0,083
GTVLU	0,008	2,091	0,037
KI	-0,040	-2,294	0,022
doba radioterapie	0,091	1,714	0,087

Tabulka 5.24: Model M2, $l(M2) = -285,719$

Krok 4

V dalším kroku budeme model dále redukovat, protože některé proměnné stále přesahují významnostní hladinu 0,05. V modelu M2 jsou to (kromě dvou variant kategoriální proměnné, které ale odstranit nemůžeme, neboť p-hodnota třetí varianty je velice nízká) proměnné *GTV* a *doba radioterapie*. Protože p-hodnoty jsou téměř stejně vysoké, můžeme se individuálně rozhodnout, kterou proměnnou vyloučit. Z praktického hlediska se dá předpokládat, že tyto dvě proměnné spolu souvisí, do modelu bychom tedy měli zahrnout tu, která je primární. *Doba radioterapie* závisí na velikosti *GTV* – objemu primárního nádoru (mimo jiné), proto v modelu ponecháme právě proměnnou *GTV*.

Takto získáme model M3 (viz Tabulka 5.26) a opět budeme zjišťovat, zda odebrání proměnné *doba radioterapie* mělo vliv na změnu modelu. Stejně jako v předchozím kroku toto budeme posuzovat pomocí změn hodnot koeficientů (viz Tabulka 5.25) a pomocí testu parciálním věrohodnostním poměrem

$$G = 2(l(M2) - l(M3)) = 2(-285,719 - (-287,103)) = 2,768,$$

statistika G má p-hodnotu rovnu 0,096, tzn. vyšší než je hladina významnosti 0,05. Změny v koeficientech (Tabulka 5.25) jsou menší než 20%, kromě změny u jedné varianty kategoriální proměnné *Dg dle MKN10* (konkrétně *dg2*).

Dospěli jsme k závěru, že odebrání proměnné *doba radioterapie* nemělo statisticky významný vliv na změnu modelu.

proměnná	změna koeficientu (%)
dg2	50,9
dg3	3,0
dg4	18,3
GTV	14,6
GTVLU	1,5
KI	0,6

Tabulka 5.25: Změny koeficientů

proměnná	koeficient $\hat{\beta}_i$	z	p-hodnota
dg2	-0,361	-0,416	0,678
dg3	-0,971	-2,832	0,005
dg4	-0,500	-1,724	0,085
GTV	0,006	2,033	0,042
GTVLU	0,008	2,199	0,028
KI	-0,041	-2,270	0,023

Tabulka 5.26: Model M3, $l(M3) = -287,103$

Všechny p-hodnoty modelu M3 uvedené v Tabulce 5.26 jsou (až na p-hodnoty dvou variant kategoriální proměnné *dg*) nižší než hladina významnosti 0,05. To znamená, že všechny proměnné jsou významné a v modelu je ponecháme. Kdyby byly některé p-hodnoty vyšší než 0,05, pokračovali bychom v postupném odebírání tak dlouho, dokud bychom nedosáhli tohoto stavu.

Krok 5

V dalším kroku budeme do modelu přidávat ty proměnné, které se při jednorozměrném log-rank testu jevíly jako statisticky nevýznamné. Ač v jednorozměrném testu byly nevýznamné, v přítomnosti jiných proměnných statisticky významné být mohou. Postupně

tedy budeme do modelu M3 přidávat proměnné: *pohlaví*, *prognóza dle NCCN*, *grading*, *dávka na frakci*, *věk* a *fyz. dávka*.

proměnná	statistika G	df	p-hodnota
pohlaví	0,023	1	0,880
prognóza dle NCCN	0,592	2	0,744
grading	3,239	3	0,356
dávka na frakci	3,267	2	0,195
věk	0,028	1	0,868
fyz. dávka	0,234	1	0,629

Tabulka 5.27: Testování modelu M3 s modelem, který vznikl přidáním jedné proměnné do M3

V Tabulce 5.27 máme uvedeny výsledky testování statistikou G. Porovnávali jsme model M3 s modelem vytvořeným z M3 přidáním jedné, původně nevýznamné, proměnné. Vidíme, že žádná z těchto proměnných není statisticky významná ani při přidání do našeho vícerozměrného modelu.

Tímto krokem sestavování Coxova modelu končí. Výsledným modelem je model M3 z Tabulky 5.26 (dále jen „model“).

5.3 Aplikace regresní diagnostiky

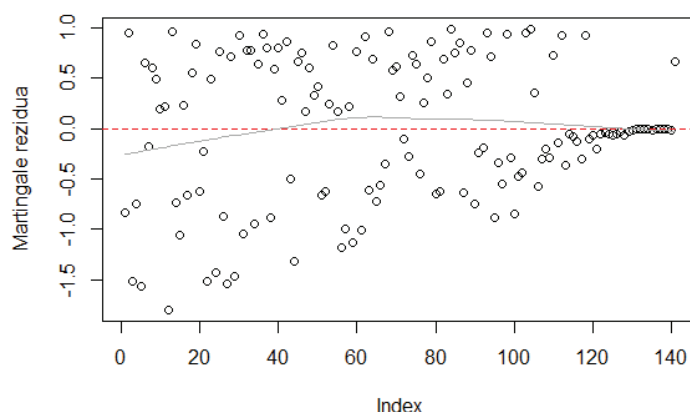
V této sekci na náš model aplikujeme regresní diagnostiku popsanou v Kapitole 4. Tím zjistíme, jestli model je nebo není adekvátní.

5.3.1 Analýza reziduí

Martingale rezidua

Velká záporná hodnota martingale rezidua znamená to, že událost ve skutečnosti nastala daleko později, než jsme na základě modelu predikovali. Naopak pokud je reziduum blízké hodnotě 1, potom událost nastala podstatně dříve, než bychom určili na základě našeho modelu. Model považujeme za vhodný, pokud martingale rezidua mají nulovou střední hodnotu.

Obrázek 5.14 znázorňuje martingale rezidua našeho modelu vykreslená proti indexu pacientů. Vidíme, že střední hodnota reziduí je přibližně rovná nule a nepozorujeme zde žádný výrazný trend. Z tohoto hlediska můžeme model považovat za vhodný.

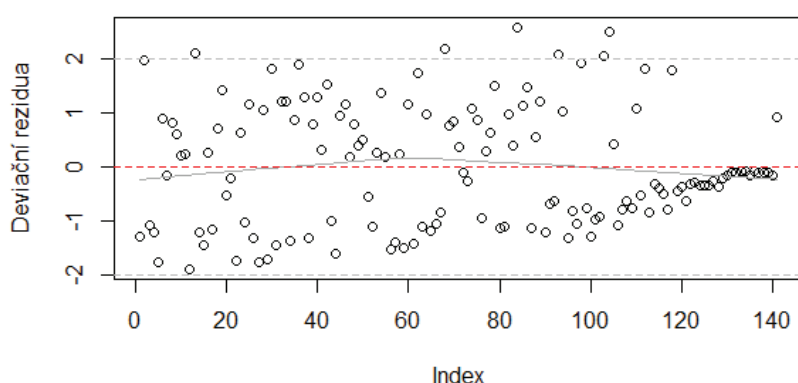


Obrázek 5.14: Martingale rezidua

Deviační rezidua

Deviační rezidua by měla být symetricky rozmístěna ve vodorovném pásu s nulovou střední hodnotou.

Na Obrázku 5.15 si můžeme prohlédnout graf deviačních reziduí vykreslených proti indexu pacientů. Vidíme, že střední hodnota je přibližně rovna nule a není patrný žádný zřetelný trend. Tato rezidua dopadla pro náš model relativně dobře.

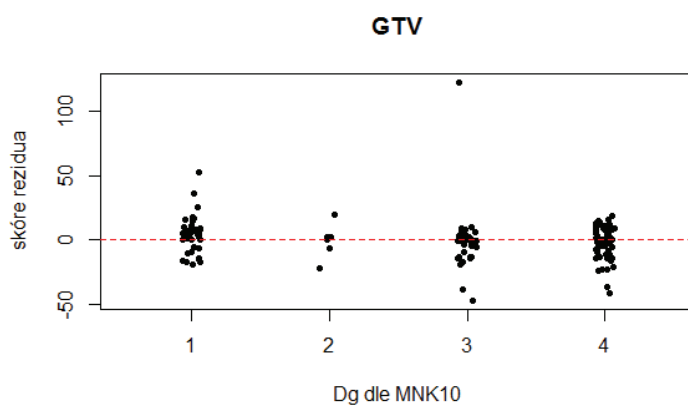


Obrázek 5.15: Deviační rezidua

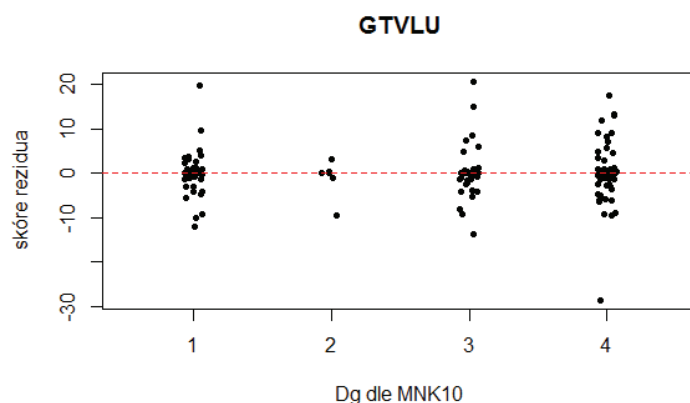
Deviační rezidua nám také pomohou odhalit odlehlá pozorování – hodnoty reziduí v těchto bodech jsou 2,581 (pacient č. 84) a 2,490 (pacient č. 104). Při bližším pohledu jsme zjistili, že tyto pacienti mají velice nízké hodnoty *GTV* a dokonce nulové hodnoty *GTVLU*, tedy bychom očekávali vyšší hodnoty *výkonnostního stavu KI*. Nicméně, není tomu tak a hodnoty *KI* jsou naopak nízké.

Skóre rezidua

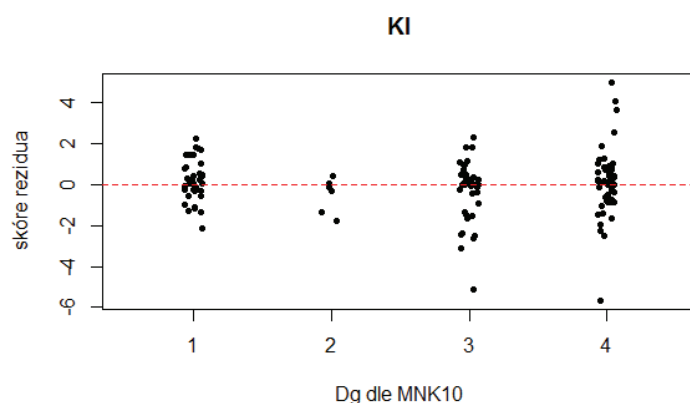
Na Obrázcích 5.16, 5.17 a 5.18 si můžeme prohlédnout grafy skórových reziduí spojitých proměnných modelu (*GTV*, *GTVLU* a *KI*) rozdělených do čtyř variant kategoriální proměnné *Dg dle MNK10*. Z těchto reziduí určíme vlivná pozorování. Z Obrázku 5.16

Obrázek 5.16: Skóre rezidua pro proměnnou *GTV*

je pro proměnnou *GTV* patrný výskyt jednoho vlivného pozorování u varianty *dg3*. Jde o pacienta s číslem 35, pro kterého je hodnota rezidua 122,575. Z Obrázku 5.17 určíme vlivná pozorování pro proměnnou *GTVLU*. Jde o jedno pozorování s hodnotou rezidua



Obrázek 5.17: Skóre rezidua pro proměnnou GTVLU



Obrázek 5.18: Skóre rezidua pro proměnnou KI

-28,572 u varianty *dg4* (pacient č. 104). Na Obrázku 5.18 jsou vidět dvě vlivná pozorování. Jedno u varianty *dg3* (hodnota rezidua: -5,138, pacient č.1), druhé u varianty *dg4* (hodnota rezidua: -5,646, pacient č.104). V Tabulce 5.28 jsou pro tři pacienty, u nichž bylo určeno vlivné pozorování, uvedeny všechny hodnoty proměnných našeho modelu.

číslo pacienta	dg	GTV	GTVLU	KI
1	3	45	25	80
35	3	20	200	75
104	4	30	0	50

Tabulka 5.28: Pacienti, u kterých bylo určeno vlivné pozorování

Pacient č. 104 má nízké hodnoty *GTV* a *GTVLU*. Při takových hodnotách bychom očekávali vysoký výkonnostní stav *KI*, ovšem zde je tomu právě naopak. Obrácená situace je u pacienta č.35, kde je vysoká hodnota *GTVLU*, ale i vysoký výkonnostní stav. U pacienta

č. 1 není na první pohled patrná příčina, proč u něj bylo identifikováno vlivné pozorování.

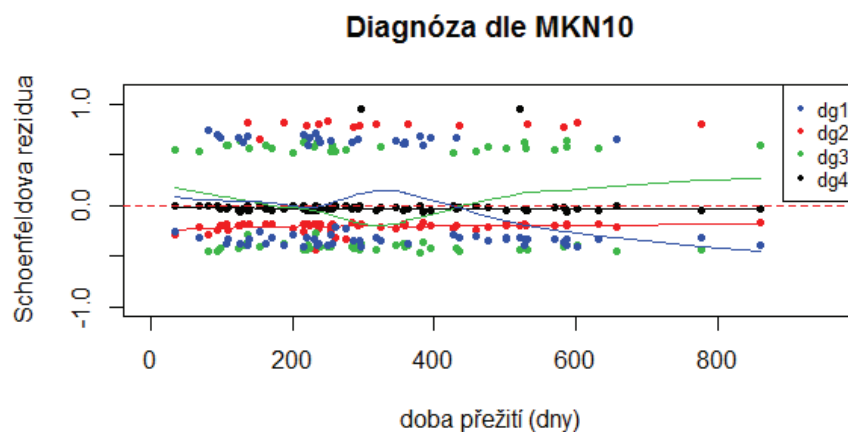
Další postup by byl následující:

Spočítáme změny v koeficientech modelu, který vznikne po vyřazení těchto tří pacientů s vlivnými pozorováními. Určíme proměnné, u kterých je změna větší než 20% a následně vyřadíme pacienty, kteří tyto změny způsobili. Tento postup není pevně daný, vždy záleží především na praktické interpretaci těchto odchylek. Pokud by to smysl mělo, pacienty s vlivnými pozorováními v modelu můžeme ponechat.

Tento postup zde provádět nebudeme, protože přesahuje rámec této diplomové práce. Bylo by však zajímavé se tomuto konceptu věnovat v budoucnu a zpřesnit tak výsledky modelování.

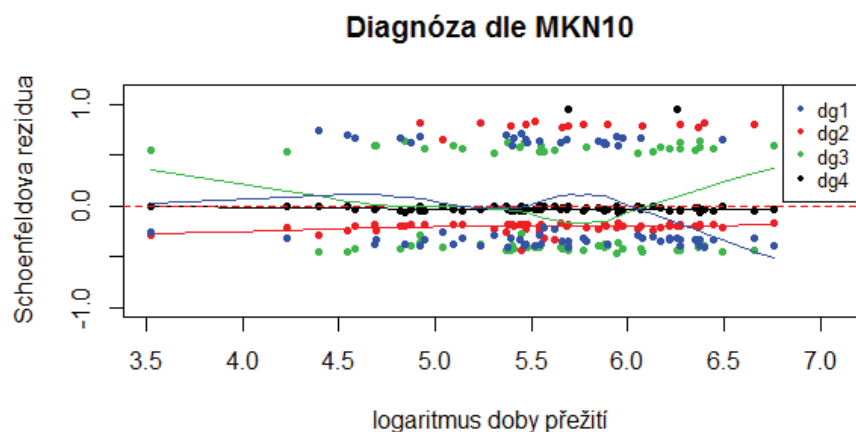
Schoenfeldova rezidua

Obrázky 5.19, 5.20 a 5.21 znázorňují Schoenfeldova rezidua našeho modelu pro čtyři varianty kategoriální proměnné *Dg dle MKN10* vykreslené postupně proti době přežití, logaritmu doby přežití a Kaplanově-Meierově odhadu. Poklesy a růsty na okrajích definičního oboru, které jsou zřejmé na Obrázcích 5.19 a 5.20 jsou způsobeny nízkým počtem pacientů pro tyto okrajové hodnoty dob přežití, resp. logaritmů dob přežití. Obecně dopadla tato Schoenfeldova rezidua pro náš model dobře, střední hodnoty jsou rovny přibližně nule.

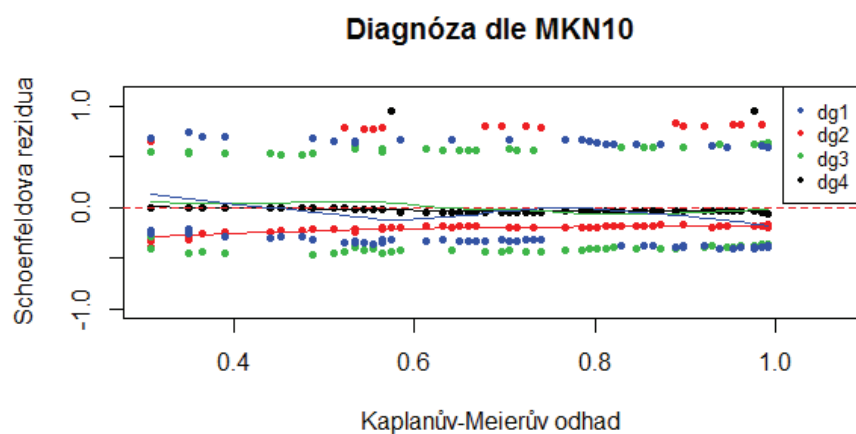


Obrázek 5.19: Schoenfeldova rezidua pro 4 varianty kategoriální proměnné *Dg dle MKN10* vykreslené proti době přežití

Na Obrázcích 5.22, 5.23 a 5.24 jsou znázorněna Schoenfeldova rezidua našeho modelu postupně pro spojité proměnné *GTV*, *GTVLU* a *KI* vykreslená vždy proti době přežití, logaritmu doby přežití a proti Kaplanově-Meierově odhadu. Na prvních dvou grafech Obrázku 5.22 (proměnná *GTV*) je patrný velice mírný stoupající trend. Pro proměnnou *GTVLU* (Obrázek 5.23) dopadla rezidua dobře a pro proměnnou *KI* je na druhém a třetím

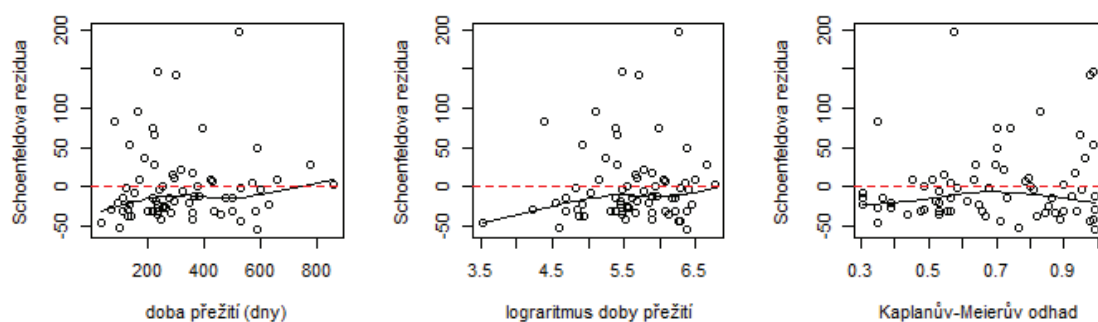


Obrázek 5.20: Schoenfeldova rezidua pro 4 varianty kategoriální proměnné Dg dle MKN10 vykreslené proti logaritmu doby přežití

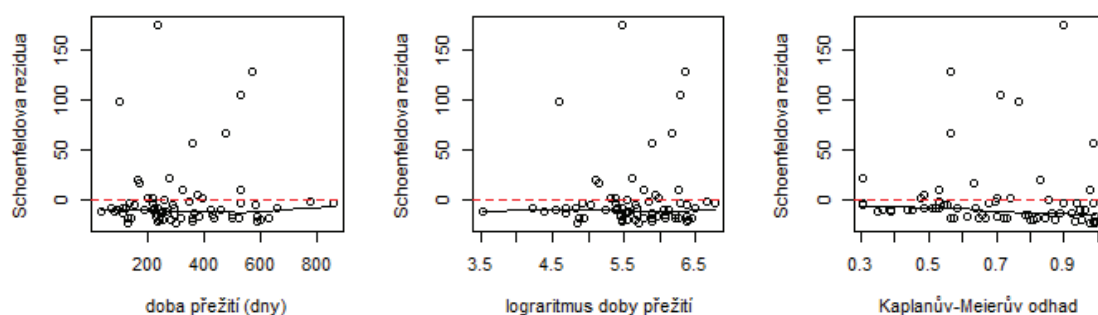


Obrázek 5.21: Schoenfeldova rezidua pro 4 varianty kategoriální proměnné Dg dle MKN10 vykreslené proti Kaplanově-Meierově odhadu

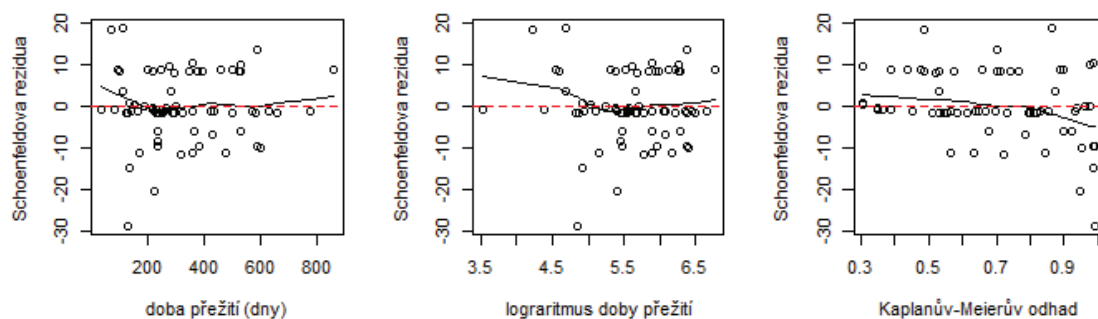
grafu Obrázku 5.24 patrný mírný klesající trend. Uvidíme, jak dopadnou pro proměnnou *KI* další testy a poté se rozhodneme, jak s ní dále naložit.



Obrázek 5.22: Schoenfeldova rezidua pro proměnnou GTV vykreslená proti době přežití, logaritmu doby přežití a proti Kaplanově-Meierově odhadu



Obrázek 5.23: Schoenfeldova rezidua pro proměnnou GTVLU vykreslená proti době přežití, logaritmu doby přežití a proti Kaplanově-Meierově odhadu



Obrázek 5.24: Schoenfeldova rezidua pro proměnnou KI vykreslená proti době přežití, logaritmu doby přežití a proti Kaplanově-Meierově odhadu

5.3.2 Hodnocení vhodnosti modelu

Pro testování celkové vhodnosti modelu jsme použili test dobré shody z balíčku `gof`. V následující Tabulce 5.29 uvádíme p-hodnoty pro jednotlivé proměnné.

proměnná	p-hodnota
dg2	0,555
dg3	0,162
dg4	0,740
GTV	0,212
GTVLU	0,839
KI	0,025

Tabulka 5.29: Test dobré shody

Vidíme, že všechny p-hodnoty kromě jedné jsou větší než 0,05. P-hodnota u *KI* říká, že bychom se měli na tuto proměnnou podívat podrobněji. Podle dalšího testování se následně rozhodneme, co s touto proměnnou provést.

5.3.3 Hodnocení proporcionality rizik

Pro hodnocení proporcionality rizik použijeme test, který jsme zavedli v Kapitole 4 na straně 28. Testování proporcionality rizik bylo provedeno v softwaru R pomocí funkce `cox.zph` z knihovny `survival`.

Na Obrázku 5.25 vidíme grafické znázornění. V prvních třech grafech (*dg2*, *dg3*, *dg4*) je patrné mírné kolísání, čtvrtý (*GTV*) a pátý (*GTVLU*) graf se jeví bez výrazného trendu a v šestém grafu (*KI*) si můžeme povšimnout, že křivka roste.

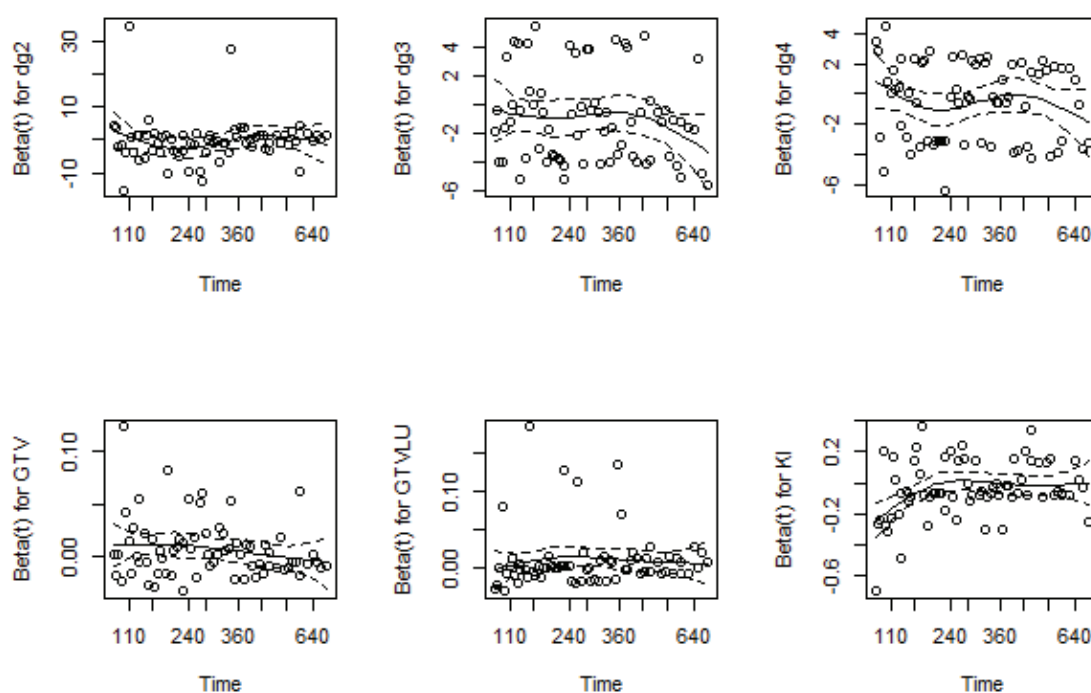
proměnná	df	$g(t) = t$		$g(t) = \ln(t)$		$g(t) = \hat{S}_{KM}(t)$	
		χ^2	p	χ^2	p	χ^2	p
dg2	1	0,071	0,790	0,009	0,923	0,048	0,826
dg3	1	2,607	0,106	1,177	0,278	1,618	0,203
dg4	1	1,363	0,243	1,905	0,167	0,697	0,404
GTV	1	2,977	0,085	2,247	0,134	2,551	0,110
GTVLU	1	0,010	0,922	0,313	0,576	0,033	0,857
KI	1	5,407	0,020	14,298	<0,001	6,657	0,010
GLOBAL	6	12,793	0,046	19,301	0,004	12,404	0,054

Tabulka 5.30: Skóre testy proporcionality rizik založené na škálovaných Schoenfeldových reziduiích (testy pomocí Pearsonova korelačního koeficientu)

V Tabulce 5.30 jsou zaznamenány výsledky testování proporcionality rizik pro tři různá škálování Schoenfeldových reziduií – hodnoty chí-kvadrát statistiky a p-hodnoty (zde označeny jako „p“). Všechny tři skóre testy vyhodnotily na významnostní hladině 0,05 proporcionalitu rizik proměnné *KI* jako nesplněnou. Toto nesplnění je dostatečně vysoké na to, aby způsobilo i porušení proporcionality rizik celého modelu (podle dvou

z těchto tří testů). Tato proměnná vykazovala problémy již při analýze reziduí. Proměnná *KI* má pravděpodobně nelineární vliv a měla by být vhodně transformována a poté by měl být vytvořen nový model.

K nalezení vhodné transformace (tzn. vyšetření linearity spojitě proměnné *KI*) může být použita například metoda využívající kvartily návrhové proměnné (*quartile design variable method*) nebo metoda frakčních polynomů (*fractional polynomials method*). Těmto metodám se věnují např. Hosmer, Lemeshow a May, 2008. Nicméně, tato problematika překračuje rozsah této diplomové práce, proto se jí zde zabývat nebudeme.



Obrázek 5.25: Testování proporcionality rizik

5.4 Interpretace výsledného modelu

Výsledný model obsahuje proměnné Dg dle $MKN10$, GTV (objem primárního nádoru), $GTVLU$ (objem uzlinové metastázy) a KI (výkonnostní stav). V Tabulce 5.31 jsou uvedeny odhadnuté koeficienty $\hat{\beta}_i$ a další související hodnoty: ($\exp(\hat{\beta}_i)$, standardní chyba, z-souřadnice a p-hodnota Waldova testu).

proměnná	koeficient $\hat{\beta}_i$	$\exp(\hat{\beta}_i)$	$SE(\hat{\beta}_i)$	z	p-hodnota
dg2	-0,361	0,697	0,869	-0,416	0,678
dg3	-0,971	0,379	0,343	-2,832	0,005
dg4	-0,500	0,606	0,290	-1,724	0,085
GTV	0,006	1,006	0,003	2,033	0,042
GTVLU	0,008	1,008	0,004	2,199	0,028
KI	-0,041	0,960	0,018	-2,270	0,023

Tabulka 5.31: Konečný Coxův model

Výsledný Coxův model proporcionálních rizik je tvaru

$$h(t) = h_0(t)e^{-0,361dg_2 - 0,971dg_3 - 0,5dg_4 + 0,006GTV + 0,008GTVLU - 0,041KI}.$$

Z Tabulky 5.31 můžeme rovnou vyčíst poměry rizik (definováno v Kapitole 3, str. 16), a to ze sloupce hodnot $\exp(\hat{\beta}_i)$. Pro spojité proměnné (GTV , $GTVLU$, KI) tyto hodnoty udávají, o kolik bude riziko úmrtí vyšší, pokud navýšíme danou hodnotu o jednotku. Pro proměnnou GTV při navýšení o 1 ml se riziko úmrtí zvyšuje o 0,6%, při zvýšení $GTVLU$ o 1 ml se riziko úmrtí zvýší o 0,8% a při zvýšení hodnoty KI o 1 bod se riziko úmrtí sníží o 4%.

Z koeficientů modelu $\hat{\beta}_i$ můžeme spočítat odhady poměrů rizik pro jednotlivé spojité proměnné pro různé změny hodnot. Odhadovaný poměr rizik například pro 10 ml nárůst hodnoty GTV vypočítáme jako $\exp(10 \cdot 0,006) = 1,062$. To znamená, že pro pacienty s hodnotou GTV o 10ml vyšší je riziko úmrtí o 6,2% větší než pro pacienty s nižší hodnotou.

Pro 10 ml nárůst proměnné $GTVLU$ spočítáme odhad poměru rizik následovně: $\exp(10 \cdot 0,008) = 1,083$, z čehož vyplývá, že pro pacienty s hodnotou $GTVLU$ o 10ml vyšší je riziko úmrtí větší o 8,3%.

Koeficient u proměnné KI – výkonnostní stav je záporný, odhadovaný poměr rizik vypočteme jako $\exp(10 \cdot (-0,041)) = 0,664$. Tedy pro pacienty s hodnotou KI o 10 bodů vyšší je riziko úmrtí o 33,6% menší. Tuto nepřímou úměrnost jsme očekávali, neboť čím vyšší je KI , tím je výkonnostní stav pacienta lepší, tudíž poměr rizik menší.

V Tabulce 5.32 jsou souhrnně uvedeny vypočtené hodnoty odhadů poměrů rizik při 10 jednotkovém navýšení jednotlivých hodnot proměnných.

proměnná	poměr rizik
GTV	1,062
GTVLU	1,083
KI	0,664

Tabulka 5.32: Odhadnuté poměry rizik pro proměnné při nárůstu o 10 jednotek

Pro kategoriální proměnné můžeme jednoduše počítat poměry rizik mezi referenční proměnnou (zde $dg1$) a ostatními variantami. Vypočteme poměr rizik například mezi $dg1$ a $dg3$: $\exp(-0,971) = 0,379$. Výsledek nám říká, že pacienti s diagnózami ze skupiny $dg3$ mají riziko úmrtí o 62,1% nižší než pacienti ze skupiny $dg1$.

Odhady poměrů můžeme vyčíst přímo z Tabulky 5.31: pro pacienty ze skupiny $dg2$ je riziko úmrtí o 30,3% menší než pro pacienty z referenční skupiny $dg1$ a pro pacienty ze skupiny $dg4$ je riziko úmrtí o 39,4% menší než riziko pacientů ze skupiny $dg1$.

Velice snadno spočítáme poměry rizik pro různé hodnoty proměnných. Mějme pacienta č.1 se zhoubným novotvarem kořene jazyka (C01), což je varianta $dg1$ proměnné Dg dle MKN10, s objemem primárního nádoru 160ml (GTV), objemem uzlinové metastázy 40ml (GTVLU) a výkonnostním stavem 60 bodů (KI). Pacient č.2 má diagnostikován zhoubný novotvar mandle (C09), což je varianta $dg3$ proměnné Dg dle MKN10, s objemem primárního nádoru 30ml (GTV), objemem uzlinové metastázy 10ml (GTVLU) a výkonnostním stavem 90 bodů (KI). Poměr rizik vypočítáme následovně:

$$\begin{aligned} HR &= \frac{h(t, \mathbf{X}_1)}{h(t, \mathbf{X}_2)} = \frac{h_0(t)e^{-0,361 \cdot 0 - 0,971 \cdot 0 - 0,5 \cdot 0 + 0,006 \cdot 160 + 0,008 \cdot 40 - 0,041 \cdot 60}}{h_0(t)e^{-0,361 \cdot 0 - 0,971 \cdot 1 - 0,5 \cdot 0 + 0,006 \cdot 30 + 0,008 \cdot 10 - 0,041 \cdot 90}} = \\ &= \frac{e^{0,96 + 0,32 - 2,46}}{e^{-0,971 + 0,18 + 0,08 - 3,69}} = \frac{e^{1,18}}{e^{-4,401}} = e^{5,581} = 265,34 \end{aligned}$$

Výsledný poměr rizik říká, že pro pacienta č.1 je riziko úmrtí přibližně 265 krát vyšší než riziko úmrtí pacienta č.2. Záměrně jsme pro demonstraci vybrali extrémní hodnoty, nicméně tímto způsobem můžeme vypočítat poměry rizik pro pacienty s jakýmkoli hodnotami proměnných.

Shrnutí

Celkově tedy podle Coxova modelu můžeme říci, že pacienti ze skupiny $dg3$ mají největší šanci na přežití v porovnání s ostatními skupinami. Tato skupina obsahuje pouze jednu diagnózu C09 – zhoubný novotvar mandle.

Podle získaných poznatků z Coxova modelu můžeme dále tvrdit, že čím vyšší bude objem primárního nádoru (GTV) anebo objem uzlinové metastázy (GTVLU), tím bude doba přežití pacienta kratší. Naopak, čím vyšší bude hodnota výkonnostního stavu (KI),

tím bude doba přežití pacienta delší.

V log-rank testu jsme spočítali, že pokud je hodnota *GTV* nižší než 100ml, potom je doba přežití pacienta statisticky významně delší. Pro proměnnou *KI* nám tento test prozradil, že výrazně větší šanci na přežití mají pacienti s hodnotami vyššími než 70 bodů. Log-rank test nám také ukázal, že i proměnná *doba radioterapie* má statisticky významný vliv na přežití, a to tak, že pokud je délka doby radioterapie méně než 40 dní, doba přežití pacientů bude delší.

Závěr

V rámci této diplomové práce jsme se seznámili se základními pojmy analýzy přežití, nejpoužívanějšími neparametrickými a semi-parametrickými metodami a také jsme si představili důležité nástroje regresní diagnostiky.

Následovala praktická část, kde jsme získané poznatky aplikovali na reálná onkologická data. Nejprve jsme využili neparametrické metody, a to log-rank test, kterým jsme hodnotili, zda proměnné mají či nemají vliv na dobu přežití. Dále jsme si výsledky prohlédli v grafech, kde jsme porovnávali Kaplanovy-Meierovy křivky přežití. Zjistili jsme, že křivky přežití se statisticky významně liší pro proměnné *GTV*, *GTV+GTVLU*, *KI* a *dobu radioterapie*.

Poté jsme aplikovali semi-parametrickou analýzu, konkrétně Coxův model proporcionálních rizik. Sestavili jsme model, který obsahoval proměnné *Dg dle MKN10*, *GTV (objem primárního nádoru)*, *GTVLU (objem uzlinové metastázy)* a *KI (výkonnostní stav)*. Ostatní proměnné vyšly jako statisticky nevýznamné. Zdůrazňeme, že vliv proměnné *věk* se překvapivě neprokázal jako statisticky významný.

Dále jsme se zabývali regresní diagnostikou, kde jsme zjistili, že sestavený model není úplně bezchybný a měl by se poupravit. Bohužel tato problematika překračovala meze této diplomové práce, ovšem bylo by zajímavé se tomuto tématu do budoucna více věnovat.

I přes tento nedostatek jsme nakonec interpretovali výsledky a došli jsme k závěru, že největší vliv na přežití pacientů má proměnná *KI – výkonnostní stav*. Nicméně, právě tato proměnná vykazovala problémy při regresní diagnostice, proto by bylo potřeba podrobit ji dalšímu analyzování.

Seznam použité literatury

- [1] Andersen, P. K., Borgan O., Gill, R. D., Keiding N., *Statistical Models Based on Counting Processes*, Springer Science & Business Media, New York, 2012, 784 s. ISBN: 1461243483
- [2] Akaike, H., *A new look at the statistical model identification*, IEEE Transactions on Automatic Control, 1974, č. 19, 716–723.
- [3] Breslow N. E., *Covariance Analysis of Censored Survival Data*, Biometrics, 1974, č. 30, No. 1, 89–99
- [4] Chakraborty, B. S., *Survival Analysis and Cox Regression for Cancer Trials*, Department of Statistics, Sardar Patel University, 2013, dostupné z: <http://www.slideshare.net/bhaswatchakraborty/part-1-survival-analysis>
- [5] Clark, T. G., Bradburn, M. J., Love, S. B., Altman, D. G., *Survival Analysis Part I: Basic concepts and first analyses*, British Journal of Cancer, 2003, č. 89, 232–238.
- [6] Collett, D., *Modelling Survival Data in Medical Research*, 2. vyd., Chapman & Hall/CRC, 2003. 391 s. ISBN 15-848-8325-1
- [7] Cox, D. R., *Regression models and life tables (with discussion)*, Journal of the Royal Statistical Society, Series B, 1972, č. 34, s. 187–220
- [8] Ding, C., Yang, Z., Lv, Z., Du, C., Xiao, H., Peng, C., Cheng, S., Xie, H., Zhou, L., Wu, J., Zheng, S., *Long non-coding RNA PVT1 is associated with tumor progression and predicts recurrence in hepatocellular carcinoma patients*, Oncology Letters 9, 2015, č. 2, s. 955–963, dostupné z: <http://dx.doi.org/10.3892/ol.2014.2730>
- [9] Fürstová, J., *Metody analýzy přežití*, European Journal for Biomedical Informatics, dostupné z: <http://www.ejbi.org/en/ejbi/article/21-cs-metody-analyzy-preziti.html/>
- [10] Grambsch P., Therneau T., *Proportional hazards tests and diagnostics based on weighted residuals*. Biometrika, 1994, č. 81, s. 515–526
- [11] Greenwood, M., *The Errors of Sampling of the Survivorship Table*, Reports on Public Health and Medical Subjects, Her Majesty's Stationery Office, London, 1926, č. 33
- [12] Hosmer, D. W., Lemeshow S., May S., *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*, Wiley Interscience, USA, 2008, 401 s., ISBN 978-0-471-75499-2

- [13] Hosmer, D.W., Lemeshow, S., *Regression Modeling of Time to Event Data*, John Wiley & Sons, Inc., New York, 1999
- [14] Holčík, J., Komenda, M. (eds.) a kol. *Matematická biologie: e-learningová učebnice* [online]., 1. vyd., Masarykova univerzita, Brno, 2015, ISBN 978-80-210-8095-9, dostupné z: <http://portal.matematickabiologie.cz/>
- [15] Kaplan, E. L., Meier, P., *Nonparametric Estimation from Incomplete Observations*, Journal of the American Statistical Association, 1958, č. 53, s. 457–481
- [16] Karnofsky, D. A., Abelmann, W.H., Craver, L.F., Burchenal, J. H., *The Use of the Nitrogen Mustards in the Palliative Treatment of Carcinoma - with Particular Reference to Bronchogenic Carcinoma*, Cancer, 1948, č. 1(4), s. 634–56
- [17] Klein, J. P., *Survival analysis: techniques for censored and truncated data*, 2. vyd., Springer, New York, 2003, 536 s., ISBN: 03-879-5399-X
- [18] Kleinbaum, D. G., Klein, M., *Survival Analysis. A Self-Learning Text*, Springer, New York, 2005, 596 s.
- [19] Lee, E. T., Wang, J. W., *Statistical methods for survival data analysis*, 3. vyd., Wiley-Interscience, New Jersey, 2003, 513 s., ISBN 04-713-6997-7
- [20] Marubini, E., M. G. Valsecchi, *Analysing survival data from clinical trials and observational studies*, Wiley, New York, 1995, 414 s., ISBN 04-719-3987-0
- [21] May, S., Hosmer D.W., *A Cautionary Note on the Use of the Grønnesby and Borgan Goodness-of-Fit Test for the Cox Proportional Hazards Model*, Lifetime Data Analysis, 2004, roč. 10, č. 3, s. 283–291, ISSN 1380-7870
- [22] NCCN Categories of Evidence and Consensus, National Comprehensive Cancer Network, dostupné z: http://www.nccn.org/professionals/physician_gls/categories_of_consensus.asp
- [23] Ng'andu, N. H., *An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model*, 1997, Statistics in medicine, roč. 16, č. 6, s. 611–626
- [24] Parzen, M., Lipsitz S. R., *A global goodness-of-fit statistic for Cox regression models*, Biometrics, 1999, č. 55, s. 580–584
- [25] Pavlík, T., *Aplikovaná analýza přežití*, 2015, dostupné z: <http://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat-aplikovana-analyza-preziti>
- [26] Rodríguez, G., *Lecture Notes on Generalized Linear Models*, 2007, dostupné z: <http://data.princeton.edu/wws509/notes/>
- [27] Zvárová, J., Malý, M., *Biomedicínská statistika III., Statistické metody v epidemiologii*, Univerzita Karlova, Praha, 2003, 505 s., ISBN 80-246-0764-6

